Description et prédiction à partir de données structurées en plusieurs tableaux : Application en épidémiologie animale.

Stéphanie Bougeard

Université Rennes 2, Haute Bretagne

Numéro _____ Année 2007

Description et prédiction à partir de données structurées en plusieurs tableaux. Application en épidémiologie animale.

Thèse présentée pour obtenir le grade de
Docteur de l'Université Rennes 2
Discipline : Statistique

par

Stéphanie Bougeard

Soutenue publiquement le 11 décembre 2007 devant le jury composé par :

M. Carbon	Professeur, Université Rennes 2	Directeur de thèse
E.M. Qannari	Professeur, *ENITIAA*, Nantes	Co-directeur de thèse
M. Hanafi	Ingénieur de recherche, *ENITIAA*, Nantes	Co-directeur de thèse
P. Cazes	Professeur, Université Dauphine, Paris	Rapporteur
G. Saporta	Professeur, *CNAM*, Paris	Rapporteur
B. Faye	Ingénieur de recherche, *CIRAD*, Montpellier	Examinateur
A. Mom	Maître de conférences, Université Rennes 2	Examinateur

Remerciements

Je tiens à remercier tout particulièrement Mostafa Qannari qui a accepté de diriger ce travail de recherche. La qualité de son encadrement, ses conseils et sa disponibilité en ont permis la réalisation. Merci de m'avoir fait confiance. Je tiens ensuite à remercier Mohamed Hanafi pour toutes les connaissances qu'il m'a transmises dans le domaine des analyses multiblocs, ainsi que pour sa grande rigueur. Son recul sur le sujet a beaucoup apporté à ce travail de recherche. Mes remerciements vont également à toute l'équipe de recherche de l'Ecole Nationale d'Ingénieurs des Techniques des Industries Agricoles et Alimentaires (*ENITIAA*, Nantes), notamment Evelyne Vigneau, Michel Semenou et Philippe Courcoux. C'est leur compétence, leur enthousiasme et leur qualité pédagogique qui m'ont donné le goût de la recherche en statistique.

Je remercie Pierre Cazes et Gilbert Saporta pour l'intérêt qu'ils ont porté à ces recherches. Merci d'avoir accepté d'être rapporteurs de cette thèse et d'y ajouter de la valeur au travers de critiques constructives. Je remercie sincèrement Bernard Faye et Alain Mom d'avoir accepté de participer au jury de thèse. Je tiens enfin à remercier Michel Carbon de m'avoir ouvert les portes de son école doctorale et de m'avoir accordé sa confiance.

Ce travail de recherche s'est déroulé au sein de l'équipe d'Epidémiologie et Bien-Etre Porcin de l'Agence Française de Sécurité Sanitaire des Aliments (*AFSSA*, Ploufragan). Je tiens tout d'abord à remercier Gilles Salvat, directeur de cet établissement, pour son accueil et l'intérêt qu'il a porté à ces travaux de recherche. Je souhaite témoigner ma reconnaissance à François Madec, responsable de cette unité, pour m'avoir encouragé à faire cette thèse, soutenu dans la réalisation de ce travail et laissé entière liberté dans mes travaux de recherche. Toutes mes amitiés vont à l'équipe d'épidémiologie de l'*AFSSA* de Ploufragan, notamment Claire Chauvin, Nicolas Rose et Christelle Fablet, pour leurs qualités humaines, leur disponibilité et tout ce qu'il m'ont appris. Ils m'ont transmis la connaissance de l'épidémiologie, de son application sur le terrain et des limites des traitements statistiques usuels appliquées à ces données. Et je remercie sincèrement l'équipe technique d'épidémiologie, Virginie Dorenlor, Florent Eono, Eric Eveno et Jean-Pierre Jolly , dont l'énergie sans limite et l'enthousiasme ont permis, entre autre, le recueil des données utilisées pour ce travail de recherche.

Merci à ma famille, à mes amis, et surtout à Michel, Juliette et Cécile ... pour tout le reste !

Table des matières

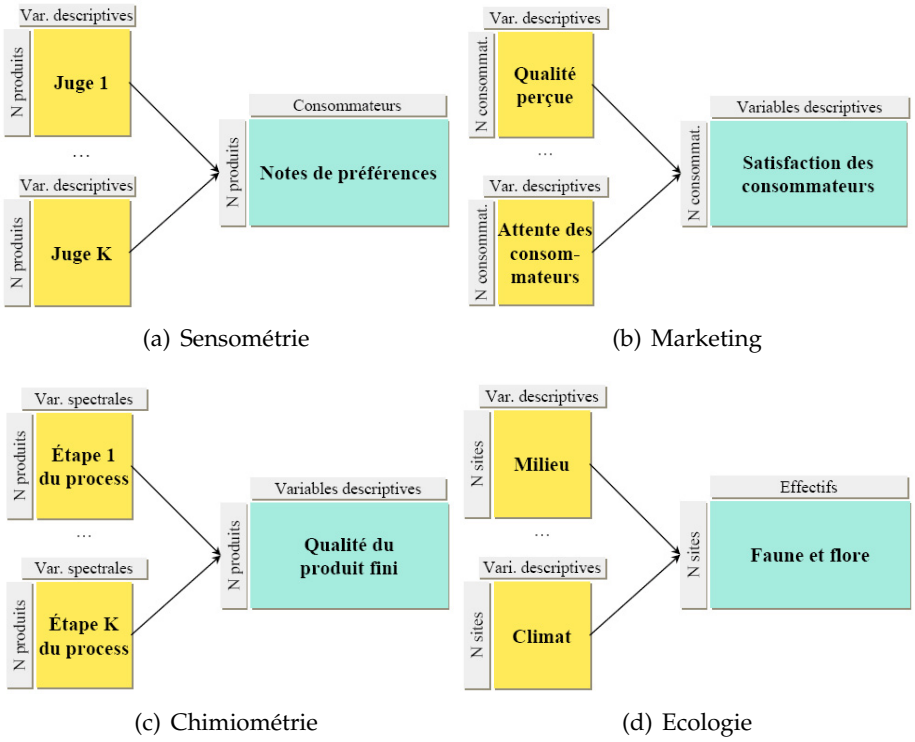
Table des figures

Liste des tableaux

Introduction

Ce travail de recherche porte sur les méthodes factorielles permettant l'analyse simultanée de plusieurs tableaux et leur application en épidémiologie animale. Les méthodes de régression multibloc, qui permettent d'orienter l'analyse d'un ou plusieurs tableaux vers l'explication d'un ou plusieurs autres, sont plus particulièrement étudiées. Ces méthodes sont principalement utilisées dans les domaines variés tels que la psychométrie, la chimiométrie, la sensométrie, l'écologie, les études de marché et les sciences sociales. Dans le domaine de la sensométrie, appliquée aux produits agro-alimentaires ou cosmétiques, les méthodes de régression multibloc peuvent servir à expliquer la préférence de consommateurs à partir de descriptions objectives fournies par plusieurs juges experts ayant évalué les mêmes produits (figure 1(a)). Une liste non exhaustive d'exemples d'utilisation est donnée par Lafosse *et al.* [1997]; Tenenhaus [1998]; Guinot *et al.* [2001]; Pagès et Tenenhaus [2001]; Tenenhaus et Vinzi [2005]; Tenenhaus *et al.* [2005a]. Toujours centré sur la satisfaction du consommateur, le domaine des études de marché utilise parfois les régressions multiblocs (figure 1(b)) [Derquenne et Hallais, 2004; Tenenhaus *et al.*, 2005b]. Dans le domaine des sciences sociales, des études ont également été présentées [Tenenhaus, 1999; Bry, 2004], expliquant l'instabilité politique de différents pays ou le coût d'un loyer, à partir de différents groupes de variables. Dans le domaine de l'écologie, l'abondance d'espèces faunistiques ou floristiques, ou l'étude de leurs pathologies, peuvent être expliquées par des groupes de variables relatives au milieu et au climat par exemple (figure 1(d)) [Hanafi et Lafosse, 2001]. Les tableaux explicatifs peuvent aussi provenir de variables relevées à différents temps. La chimiométrie est certainement le domaine qui compte le plus d'exemples d'utilisation des méthodes de régression multibloc (figure 1(c)). De nombreuses variables sont mesurées à différentes étapes d'un processus industriel ; elles sont utilisées pour expliquer et prédire la qualité finale de produits (agro-alimentaires, pharmaceutiques, ...) [Wold *et al.*, 1987; Bro, 1996; Wold *et al.*, 1996; Westerhuis et Coenegracht, 1997; Berglund et Wold, 1999; Qin *et al.*, 2001; Kourti, 2003; Vivien *et al.*, 2005; Höskuldsson et Svinning, 2006].

Les objectifs de ce travail de recherche sont à la fois méthodologiques et appliqués. Plusieurs méthodes statistiques existantes sont tout d'abord présentées et reliées dans un cadre unifié, relevant soit de critères à maximiser comparables, soit d'un continuum général reliant l'ensemble des méthodes présentées. De nouvelles méthodes statistiques sont également proposées et situées dans le cadre général en focalisant sur leur pertinence pour le traitement des données en épidémiologie



(a) Sensométrie (b) Marketing

(c) Chimiométrie (d) Ecologie

FIG. 1 – Exemple de données structurées en $(K+1)$ tableaux.

animale. Les contraintes de ces nouvelles méthodes sont issues des limites des traitements statistiques actuels des données d'épidémiologie animale. Un des principaux objectifs de ce travail de recherche est de contribuer à la réflexion sur la sensibilité à la multicolinéarité des méthodes de régressions multiblocs. Le comportement de ces méthodes, ainsi que des continuums développés, est illustré sur des applications provenant de données réelles d'épidémiologie. Les méthodes statistiques permettant de lier deux tableaux sont tout d'abord présentées. Une extension de ces méthodes pour le traitement de $(K+1)$ tableaux est ensuite proposée. L'utilisateur peut ainsi facilement accéder aux différentes méthodes, selon la nature de ses données. Des résumés, sous forme de tableaux de synthèse, sont proposés pour en faciliter la compréhension et l'accès. L'ensemble des méthodes présentées est programmé dans le cadre du logiciel Matlab.

Ce travail est organisé en trois parties. La partie I présente le contexte du travail de recherche issu des contraintes statistiques associées aux données d'épidémiologie animale. Au sein de cette partie, le chapitre 1 présente l'épidémiologie dans le cadre de l'étude des maladies et des problèmes de sécurité sanitaire des aliments d'origine animale. Nombre d'enquêtes d'épidémiologie animale visent à mettre en relief les facteurs de risque de maladies au déterminisme souvent pluri-factoriel. Les données collectées sont nombreuses et de format varié : les individus ont une structure hiérarchisée, les variables explicatives sont nombreuses, multicorrélées,

organisées en groupes ayant un sens biologique et orientées vers l'explication de plusieurs variables à expliquer. Le chapitre 2 expose une revue bibliographique sur les traitements statistiques usuels appliqués aux données d'épidémiologie animale analytique. Les analyses relatives à l'utilisation de l'analyse de données sont plus particulièrement détaillées. Le travail de recherche est centré sur l'exploration de certaines problématiques majeures du traitement statistique des données d'épidémiologie animale, permettant de décrire et modéliser le lien entre les variables explicatives et les variables à expliquer, en considérant plusieurs variables à expliquer, des variables explicatives organisées en groupes. Le choix qui est fait pour répondre à ces problématiques est basé sur des méthodes factorielles appropriées.

La partie II est axée sur l'étude des méthodes qui décrivent et relient deux tableaux de variables X et Y. Le chapitre 3 présente les principales méthodes répondant à cette problématique, à savoir la régression sur composantes d'*ACP*, la *latent root regression*, l'analyse canonique, la régression *PLS* et l'analyse en composantes principales sur variables instrumentales. Une vision synthétique de ces méthodes est présentée au travers de critères à maximiser uniformisés, associés à différentes contraintes de normes ou de déflation. Le chapitre 4 propose tout d'abord un continuum permettant de relier toutes les méthodes présentées précédemment. Les continuums les plus connus, régression *ridge*, *principal covariate regression*, extension de la méthode *continuum regression* et analyse canonique *ridge*, sont replacés comme cas particuliers de ce continuum. Par la suite, deux nouveaux continuums, adaptés au traitement des données d'épidémiologie animale analytique sont proposés. Le premier, appelé continuum *LRR* et basé sur une extension de la *latent root regression*, détermine le poids du tableau Y dans la construction des composantes, contraintes d'être dans l'espace des variables X. Le second établit le lien entre une méthode pouvant être instable mais explicative de Y, l'analyse en composantes principales sur variables instrumentales, et une méthode plus stable mais moins explicative de Y, la régression *PLS*. Une application au traitement de données d'épidémiologie animale organisées en deux tableaux est proposée dans le chapitre 5. Les deux continuums développés, la méthode *principal covariate regression*, ainsi que leurs cas particuliers sont étudiés et comparés. Les propriétés de stabilité, de capacité prédictive et l'évolution des cartes factorielles de ces continuums sont illustrées. Les performances prédictives des continuums et de leurs cas particuliers sont comparées sur la base d'une validation croisée.

La partie III présente des méthodes qui décrivent K tableaux X_k ($k = 1, \ldots, K$) et qui sont orientées vers l'explication d'un tableau Y. Les méthodes du chapitre 6 sont des extensions multiblocs des méthodes développées dans le chapitre 3. Différentes méthodes multiblocs basées sur des critères à maximiser clairs et répondant aux objectifs de traitement des données d'épidémiologie animale sont exposées. Elles présentent les avantages de permettre à la fois des représentations factorielles de l'ensemble des variables explicatives et à expliquer, ainsi qu'une modélisation associée. L'avantage de prendre en compte la structure en blocs des variables explicatives est tout d'abord d'équilibrer le poids des blocs dans l'explication des variables Y, mais aussi d'apporter de nouvelles réponses aux épidémiologistes comme la mesure de l'influence des blocs dans l'explication de la maladie. Les méthodes multiblocs

présentées tout d'abord s'apparentent à l'analyse canonique, mais leur sensibilité à la multicolinéarité au sein des différents tableaux les rend peu adaptées au traitement de données souvent multicolinéaires. Les méthodes multiblocs s'apparentant à l'analyse en composantes principales sur variables instrumentales limitent cette sensibilité sans toutefois s'en affranchir totalement. Des méthodes moins sensibles à la multicolinéarité sont reprises (*PLS* multibloc) et proposées (*latent root regression* multibloc). Une vision synthétique de ces méthodes est présentée au travers de critères à maximiser uniformisés, associés à différentes contraintes de normes ou de déflation. Le chapitre 7 propose plusieurs continuums reliant les méthodes présentées dans le chapitre 6 afin de mieux comprendre leur fonctionnement. Certains continuums modifient le critère à maximiser, plus ou moins orienté vers l'explication des variables des différents blocs. D'autres permettent d'appréhender l'influence des différentes contraintes de normes sur les résultats obtenus. Le chapitre 8 présente une application à des données d'épidémiologie animale des méthodes et continuums les plus intéressants présentés dans les chapitres 6 et 7. La description des données structurées en blocs est tout d'abord étudiée au travers de l'inertie des différents tableaux expliquée par les composantes, ainsi que des cartes factorielles associées. Les performances prédictives des continuums multiblocs et de leurs cas particuliers sont comparées sur la base d'une validation croisée.

Première partie

Traitement statistique des données d'épidémiologie animale

Chapitre 1

Structure des données d'épidémiologie animale

1.1 Notions d'épidémiologie

1.1.1 Définition générale

Pour les vétérinaires, l'épidémiologie est l'étude des maladies et des phénomènes de santé dans une population animale [Toma *et al.*, 1996]. Les phénomènes sanitaires considérés sont aujourd'hui souvent d'origine multi-factorielle et résultent d'interactions entre un ou plusieurs agents pathogènes, une population et le milieu dans lequel vit celle-ci. Les propriétés de chacun de ces trois éléments conditionnent l'épidémiologie de la maladie et sont résumées par la figure 1.1.

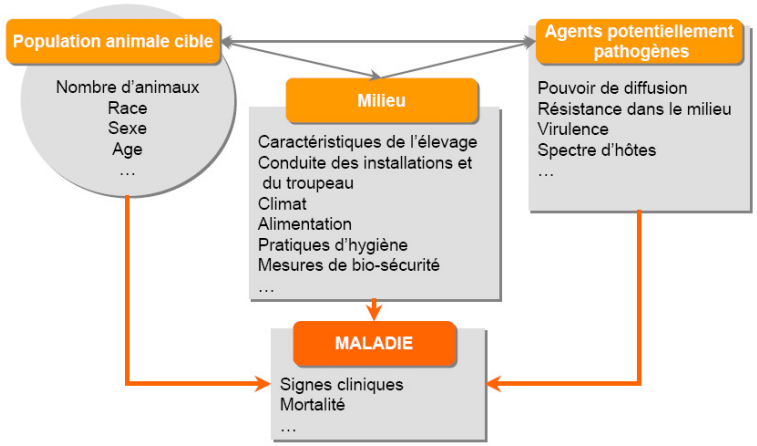


Fig. 1.1 – Cadre général simplifié de l'épidémiologie animale, d'après Toma *et al.* [1996].

1.1.2 Les différentes cibles de l'épidémiologie vétérinaire

Dans le domaine vétérinaire, l'épidémiologie réunit différentes étapes d'une démarche globale qui vise à lutter contre les maladies [Toma *et al.*, 1996]. Les données sont obtenues le plus souvent au travers d'enquêtes conduites en élevages et/ou dans des laboratoires de diagnostic. Pour une maladie donnée, la première étape est l'épidémiologie descriptive. Elle consiste à caractériser l'évolution dans le temps et dans l'espace d'une maladie dans une population. Ce type d'enquête peut être complété par des études expérimentales, en station ou au laboratoire, qui consistent en l'étude d'un facteur présumé causal dans une situation expérimentale où tous les autres facteurs sont contrôlés. Elle se fait soit au laboratoire, soit sur le terrain sur une population limitée. Les résultats fournissent des éléments essentiels à l'orientation de la seconde et principale étape, l'étude analytique, en posant des hypothèses sur les facteurs de risque potentiels [Bouyer *et al.*, 1995]. De manière plus spécifique, la deuxième étape a pour objectif de connaître les mécanismes de développement de la maladie. Si la maladie est transmissible et de déterminisme mono-factoriel, elle doit permettre de connaître la nature de l'agent pathogène, ses sources, ses cibles et son mode de transmission. Si la maladie relève d'un déterminisme pluri-factoriel, elle doit permettre de déterminer les facteurs associés à l'apparition et au développement de la maladie et de démontrer une relation de cause à effet entre les facteurs supposés de risque et la maladie. Dans le cadre des enquêtes d'épidémiologie vétérinaire sur ce type de pathologies complexes, ce sont les facteurs d'expression de la maladie qui sont recherchés. Deux autres étapes, basées sur le même procédé d'enquête, permettent d'exploiter les résultats issus de l'enquête analytique. L'épidémiologie opérationnelle correspond à la mise à profit des connaissances obtenues pour la conception et l'application de mesures de lutte contre la maladie visée. L'épidémiologie évaluative permet enfin d'évaluer les résultats du programme de lutte. Ces différentes étapes sont résumées par la figure 1.2.

1.1.3 L'épidémiologie analytique

L'épidémiologie analytique est l'étude des causes apparentes et des évènements directement ou indirectement associés à un phénomène de santé [Toma *et al.*, 1991]. Ce type d'étude vise à mesurer l'intensité de la liaison entre des facteurs étudiés et la maladie, et ainsi de déterminer des facteurs de risque associés au développement de la maladie. La démarche se fonde en général sur la réalisation d'enquêtes nécessitant la constitution d'un groupe témoin ne présentant pas la maladie étudiée, afin d'effectuer la comparaison indispensable à l'interprétation. Schématiquement, l'étude du rôle d'un facteur ou d'une combinaison de facteurs de risque supposés passe par l'un des deux protocoles : exposé-non exposé ou cas-témoin.

L'enquête exposé-non exposé est une étude analytique longitudinale prospective dont le principe est de comparer l'incidence (ou fréquence de survenue) d'une maladie chez des individus exposés, à celle chez des individus non exposés, pris comme témoins (figure 1.3). Tous les individus sont indemnes de la maladie au début de l'observation et diversement exposés aux facteurs de risque. La durée de l'enquête varie selon la nature de la maladie et le temps de latence entre l'exposition et la

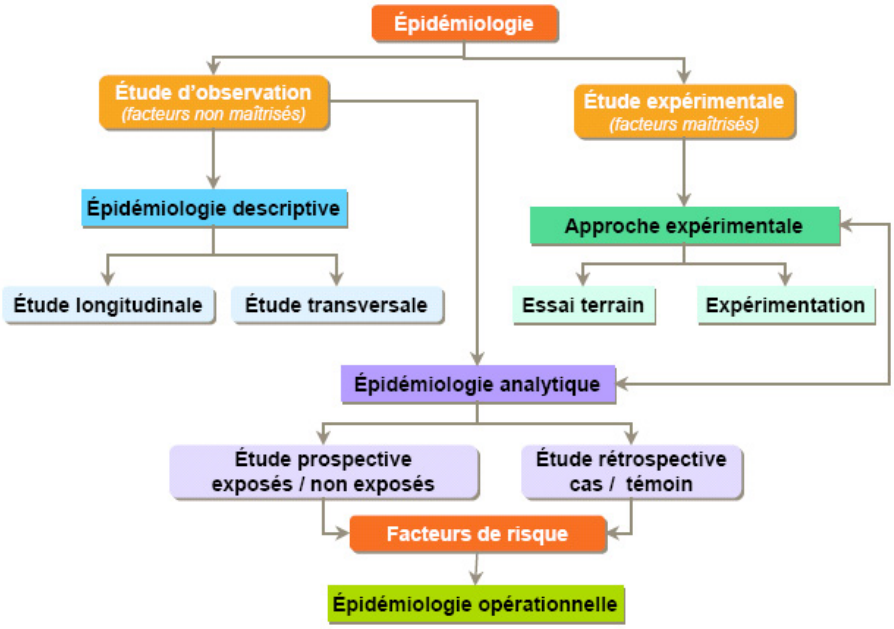


FIG. 1.2 – Principales enquêtes en épidémiologie vétérinaire.

maladie. Elle est particulièrement indiquée pour une maladie fréquente comportant un délai court entre l'exposition au facteur de risque et l'apparition de la maladie et pour laquelle le ou les facteurs étudiés sont supposés avoir un rôle causal important.

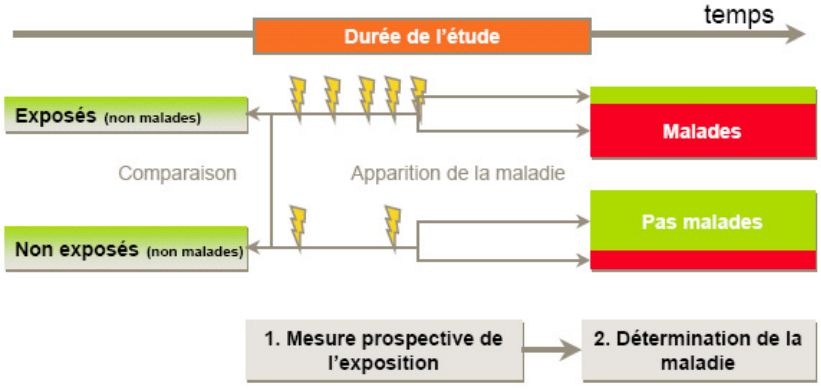


FIG. 1.3 – Enquête exposé-non exposé.

L'enquête cas-témoin est une étude analytique rétrospective, consistant à comparer un groupe de sujets malades (cas) et un groupe de sujets indemnes (témoins), grâce à la récupération d'informations sur l'exposition antérieure à des facteurs de risque [Toma *et al.*, 1991], comme illustré par la figure 1.4. Les sujets sont inclus dans l'étude au moment de la survenue de la maladie. Le groupe témoin doit être comparable au groupe cas sur le plan des caractéristiques descriptives (âge, sexe, ...) afin de permettre la comparaison entre les deux groupes. Ce type d'étude est

adapté aux maladies rares à long délai d'incubation. Il permet l'étude simultanée de plusieurs facteurs de risque car il est possible de récolter des informations sur de nombreuses variables d'exposition.

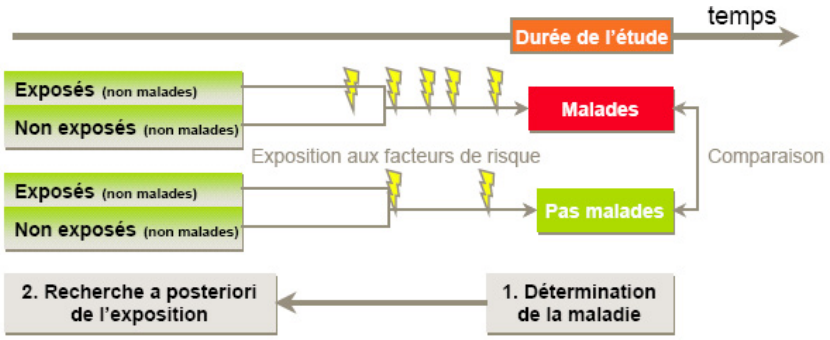


FIG. 1.4 – Enquête cas-témoin.

Ces deux types d'enquêtes analytiques sont présentés de façon schématique. Dans le cadre usuel des enquêtes d'épidémiologie vétérinaire, les enquêtes menées peuvent combiner, par exemple, un questionnaire (rétroactif) administré à l'éleveur sur les pratiques en élevage, ainsi qu'un suivi prospectif des animaux.

1.2 Quantification des causes de la maladie par les facteurs de risque

1.2.1 Définition d'un facteur de risque

Le risque est la probabilité pour un sujet non malade de devenir malade au cours d'une période fixée. Le facteur de risque est un facteur associé à l'augmentation de la probabilité d'apparition ou de développement d'une maladie. Il est statistiquement associé à une maladie et intervient en amont de la déclaration de celle-ci. Le fait de travailler sur des populations rend nécessaire la quantification des phénomènes. Seules les études expérimentales permettent d'établir les preuves de causalité. Dans les enquêtes analytiques, des hypothèses sont testées et aboutissent à des présomptions de causalité. Il existe deux acceptions du facteur de risque. On parle soit d'indicateur de risque s'il n'y a pas de causalité à l'association statistique entre le facteur et la maladie. On parle de facteur de risque s'il y a une causalité probable ou établie à l'association statistique entre le facteur et la maladie. Pour confirmer les présomptions de causalité, la relation entre risque et facteur de risque doit vérifier les critères de Bradford-Hill (revus par Evans [1978]) : association statistique forte, relation dose-effet (quand le facteur de risque est quantitatif ou ordinal), pas d'ambiguïté sur la chronologie (antériorité de l'exposition par rapport à la maladie), constance des résultats dans diverses études, plausibilité de l'hypothèse, cohérence des résultats et spécificité de l'association. Pour une discussion plus complète sur le sujet, se référer à Madec et Tillon [1988]; Madec et Fourichon [1990].

1.2.2 Quantification par le risque relatif ou l'*odds ratio*

L'intensité de la relation entre facteur et maladie est évaluée d'après la valeur du risque relatif (*RR*). Le risque relatif est le rapport du taux d'incidence dans le groupe exposé rapporté au taux d'incidence dans le groupe non exposé. Ce risque n'est en fait directement calculable que dans les enquêtes exposés-non exposés, où l'incidence de la maladie de l'échantillon est le reflet de celle de la population. Dans une enquête cas-témoin, l'incidence de la maladie n'est pas connue car les effectifs des deux groupes, cas et témoin, ne la reflètent pas. Dans une enquête cas-témoin, la mesure de la liaison entre chaque facteur étudié et la maladie se fait par le calcul d'un *odds ratio* (*OR*). Quand la maladie est rare, le risque relatif peut être approximé par l'*odds ratio* [Bouyer *et al.*, 1995]. Quand la maladie n'est pas rare, il est possible de déduire le risque relatif de l'*odds ratio* [Beaudeau et Fourichon, 1998].

Le risque relatif est la mesure du rôle du facteur étudié. Il peut prendre n'importe quelle valeur positive. L'indépendance entre le facteur de risque potentiel et la maladie est vérifiée pour une valeur du risque relatif égale à un. Une valeur de risque relatif éloignée de un représente un degré élevé d'association entre le risque potentiel et la maladie. Lorsque le risque relatif est inférieur à un, avec un non compris dans l'intervalle de confiance du risque relatif, le fait qu'un individu soit malade est moins probable pour un individu exposé que pour un individu non exposé ; le facteur de risque a un effet protecteur vis à vis de la maladie. Lorsque le risque relatif est supérieur à un, avec un non compris dans l'intervalle de confiance, le fait qu'un individu soit malade est moins probable pour un individu non exposé que pour un individu exposé ; le facteur est donc à risque vis à vis de la maladie. Une valeur de risque relatif égale à cinq, par exemple, signifie que le risque d'apparition de la maladie est cinq fois plus grand dans le groupe exposé au facteur que dans le groupe non exposé. L'*odds ratio* s'interprète de la même façon que le risque relatif. Comme l'*odds ratio* est calculable et interprétable pour tous les types d'enquête, c'est lui qui sera utilisé par la suite.

1.2.3 Lien entre l'*odds ratio* et les coefficients de régression

Dans le cas où la maladie est caractérisée par une variable dichotomique ($y = 1$ indique la présence de la maladie et $y = 0$ son absence) et n'est expliquée que par une variable d'exposition, la distribution observée des proportions des individus *cas* en fonction de la variable explicative est estimée par une courbe dont la forme correspond à la fonction logistique. Dans le cas où la variable d'exposition est dichotomique elle-aussi, il est aisément démontré que l'*odds ratio* est égal à l'exponentielle du coefficient de régression : $OR = e^{\beta}$ [Agresti, 2002, p. 124]. Dans le cas où la variable explicative est quantitative, la définition de l'*odds ratio* est la même mais correspond à l'augmentation du risque d'être malade quand cette variable augmente d'une unité [Agresti, 2002, p. 166]. On mesure ici la difficulté d'interprétation liée aux variables quantitatives. Ceci conduit à préconiser le recodage des variables quantitatives, sous la forme d'une mise en classe d'intervalles. Lorsque la maladie est caractérisée par une variable dichotomique expliquée par plusieurs variables d'exposition, l'*odds ratio* associé à chaque variable explicative correspond à l'exponentielle du coefficient

de régression associé à cette variable [Agresti, 2002, p. 183].

Lorsque la maladie est décrite par une variable quantitative, l'utilisation de la régression logistique n'est plus possible. Ce sont les valeurs des coefficients de la régression qui quantifient l'intensité des liens entre la maladie et les variables explicatives. De la même façon, les *odds ratio* sont calculés à partir de l'exponentielle des coefficients de régression. Dans le cas où la description de la maladie est faite par plusieurs variables, des *odds ratio* sont calculés pour associer chaque variable explicative à chacune des variables à expliquer. Cependant il arrive que, pour une même variable explicative, des *odds ratio* ayant des interprétations différentes soient associés aux variables à expliquer. Dans ce cas, l'interprétation devient difficile faute d'une vision synthétique.

1.3 Structure des données d'épidémiologie animale

1.3.1 Organisation pratique des enquêtes

Les études d'épidémiologie analytique dans le domaine animal sont réalisées au travers d'enquêtes menées en élevage, à l'abattoir et/ou au laboratoire. Une partie des données collectées provient d'un questionnaire administré à l'éleveur portant sur l'ensemble de ses pratiques d'élevage. Toutes ces informations sont, dans la mesure du possible, objectivées par des mesures réalisées dans l'élevage. Des données biologiques sont aussi collectées sur les animaux, au cours d'une ou plusieurs visites : poids, portage de différents agents infectieux (résultats issus d'une prise de sang, d'un lavage trachéo-bronchique, d'un prélèvement nasal, ...), température et autres signes cliniques. Un dernier ensemble de données est éventuellement collecté à l'abattoir sur les animaux suivis au préalable. Il consiste en des prélèvements et des notations de l'état de divers organes (foie, rate, ganglions, poumons, ...), ainsi que des analyses sur le portage de différents agents infectieux par ces organes.

Les enquêtes en élevage comprennent différents groupes, dont certains peuvent être communs d'une étude à une autre. Les groupes majeurs sont : les caractéristiques de la ferme (taille de l'élevage, performances zootechniques, autres productions animales, ...), la conduite d'élevage (taux de renouvellement, technique de reproduction, nombre d'animaux par portée, nombre de bandes d'animaux, ...), l'habitat des animaux (enregistrements bio-climatiques, ventilation, isolation, chauffage, ...), l'alimentation et l'abreuvement des animaux (compositions alimentaires, mode de distribution, nombre de mangeoires, origine des aliments, ...), l'état sanitaire du troupeau (dosages sérologiques, pesées, maladies chroniques, taux de réformes, vaccinations, traitements antibiotiques, ...), les pratiques d'hygiène (protocoles de nettoyage et de désinfection) et les mesures de bio-sécurité (mesures sanitaires de l'éleveur et des visiteurs, équarrissage, ...).

1.3.2 Exemple d'enquêtes d'épidémiologie animale

Le laboratoire de l'*AFSSA* (=Agence Française de Sécurité Sanitaire des Aliments) de Ploufragan est l'un des douze laboratoires de l'*AFSSA*. Il a pour mission

de contribuer à l'amélioration de la santé et du bien-être des animaux dans les productions avicoles, cunicoles et porcines, ainsi qu'à la qualité sanitaire des aliments qui en sont issus. Les travaux menés au sein des unités d'Epidémiologie consistent principalement en des enquêtes épidémiologiques analytiques. Elles visent à déterminer les facteurs de risque d'une maladie ou d'un phénomène de santé publique en relation avec les productions animales précitées.

Les données issues de cinq enquêtes d'épidémiologie analytique menées à l'*AFSSA*, vont servir d'exemple dans ce document pour illustrer la structure classique des données d'épidémiologie animale. Elles servent d'appui pour les applications des méthodes développées dans ce travail de recherche. Comme vu dans le paragraphe 1.1.3, le type général d'enquête analytique, exposé-non exposé ou cas-témoin, conditionne le mode de collecte des données, prospectif ou rétrospectif, mais ne change ni l'objectif de l'étude ni la structure des données. Les références bibliographiques indiquées dans le tableau 1.1 fournissent les détails du mode de collecte des données ainsi que les traitements statistiques usuels utilisés pour la détermination des facteurs de risque.

Enquête	**Type**	**Objectifs**	**Références bibliographiques**
SALMO porc	Exp.-Non exp.	Facteurs de risque de l'excrétion de *Salmonella* dans les élevages de porcs	Beloeil *et al.* [2003, 2004a,b]
MAP pietrain porc	Exp.-Non exp.	Facteurs de risque de la maladie de l'amaigrissement du porcelet (=*MAP*)	Rose *et al.* [2004]
EEL lapin	Cas-Témoin	Facteurs de risque de l'entérocolite épizootique de lapin (=*EEL*)	Klein [2002]
MAP élevage porc	Cas-Témoin	Facteurs de risque de la *MAP* au niveau de l'élevage	Rose *et al.* [2003a]; Venot [2003]
ANTIBIO dinde	Cas-Témoin	Facteurs de risque de la consommation d'antibiotiques en élevage de dinde	Chauvin *et al.* [2005]

Tab. 1.1 – Type et objectif des enquêtes d'épidémiologie animale menées à l'*AFSSA*, illustrés par les profils de cinq enquêtes.

Le nombre d'individus d'une enquête analytique est conditionné par la prévalence (*i.e.* pourcentage d'individus infectés ou présentant le phénomène de santé étudié dans la population) de la maladie ou du phénomène de santé étudié, afin que l'échantillon étudié soit représentatif de la population dont il est issu. Pour parfaire cette représentativité, des tirages au sort stratifiés sur certains critères d'importance (la taille de l'élevage par exemple) peuvent aussi être réalisés. L'unité est soit l'animal, soit le lot (*i.e.* groupe de volailles nées, élevées et abattues en même temps) ou la bande (*i.e.* groupe de porcs nés, élevés et abattus en même temps), soit l'élevage selon l'objectif de l'étude. Lorsque l'unité est l'animal, il est important de prendre en compte plusieurs effets emboîtés : celui de l'animal dans sa case (*i.e.* groupe d'une dizaine de porcs, parfois issus d'une même portée, élevés ensemble), de cette case dans la bande, et de la bande ou du lot dans l'élevage. Le nombre de variables X explicatives données dans le tableau 1.2 correspond aux variables interprétables

et sélectionnées au préalable pour leur lien avec la(es) variable(s) à expliquer. Ces variables explicatives sont un mélange de variables qualitatives et quantitatives organisées en plusieurs groupes. Il peut y avoir quelques données manquantes dans certaines études (de l'ordre de 10%). La maladie est souvent complexe et décrite par plusieurs variables à expliquer regroupées dans un tableau Y (taux de mortalité, signes cliniques observés en élevage et à l'abattoir, ...).

Enquête	**Nb individus**	**Nb. var.** Y	**Nb. var.** X	**Nb groupes** X_k	**Type de var.** X
SALMO porc	105 élevages	7	65	3	80% quali., 20% quanti.
MAP pietrain porc	840 animaux	5	34	4	35% quali., 65% quanti.
EEL lapin	96 élevages	4	40	7	58% quali., 42% quanti.
MAP élevage porc	159 élevages	1	43	7	42% quali., 58% quanti.
ANTIBIO dinde	659 lots	3	37	7	13% quali., 87% quanti.

TAB. 1.2 – Caractéristiques générales des données d'épidémiologie animale de l'*AFSSA* de Ploufragan, illustrées par le profil de cinq enquêtes.

1.3.3 Caractéristiques générales des données d'épidémiologie animale

Les données d'épidémiologie animale sont de nature complexe, à la fois du point de vue des individus dont la structure est multi-emboîtée, du point de vue des variables explicatives qui sont de nature mixte, organisées en groupes, contenant parfois des données manquantes et multi-corrélées, et enfin du point de vue des variables à expliquer qui sont multiples et elles-aussi de nature mixte. Certaines variables explicatives sont collectées plusieurs fois au cours de différentes visites lors de l'enquête. La maladie ou le phénomène de santé peuvent aussi être caractérisés par leur évolution au cours du temps. Ces structures temporelles ne sont pas évoquées par la suite. Nous considérons dans tous les cas que nous disposons d'une structure multibloc, les variables mesurées au cours du temps pouvant constituer des blocs spécifiques. Un exemple de la structure la plus usuelle de ces données est résumé par la figure 1.5.

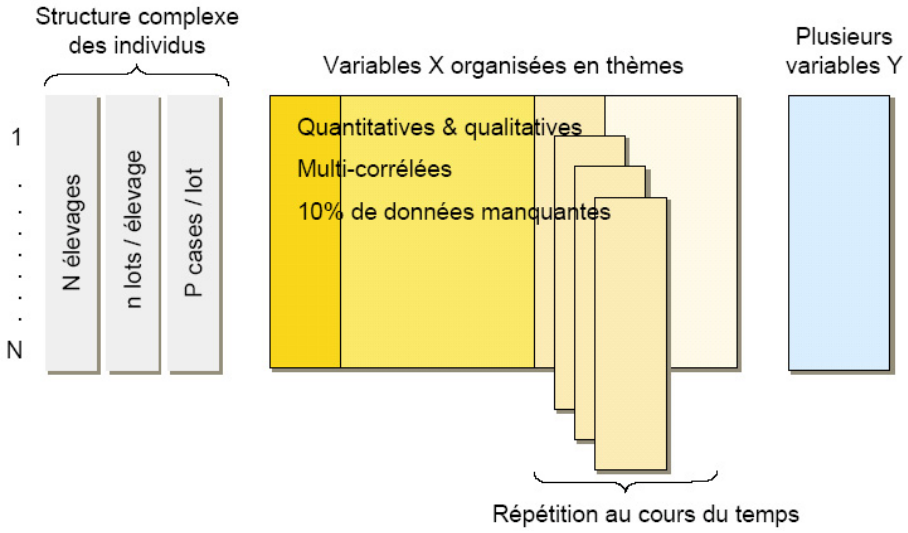


FIG. 1.5 – Illustration de la structure usuelle des données d'épidémiologie animale.

Chapitre 2

Problématique et traitement statistique en épidémiologie animale

2.1 Traitement statistique en épidémiologie animale

Seules les méthodes de traitement statistique des enquêtes épidémiologiques analytique, qu'elles soient de type cas-témoin ou exposé-non exposé, sont détaillées ici. Comme leur objectif, ainsi que la structure de leurs données sont similaires, les traitements statistiques appliqués à ces deux types d'enquête sont les mêmes. La description des traitements statistiques appliqués aux enquêtes analytiques est principalement vue au travers d'une revue bibliographique du journal *Preventive Veterinary Medicine* (=*PVM*), référence pratiquement incontournable pour le traitement de ce type d'enquêtes depuis 1982 (2427 articles parus).

2.1.1 Utilisation classique de la régression

L'objectif d'une enquête d'épidémiologie analytique est de définir les facteurs de risque d'une maladie ou d'un problème de santé publique vétérinaire. Le traitement statistique des données est donc classiquement basé sur une régression, qui permet d'expliquer la maladie y par les variables explicatives X présumées influentes.

Utilisation des modèles linéaires généralisés

Les méthodes de régression utilisées pour le traitement des données d'épidémiologie animale sont des cas particuliers du modèle linéaire généralisé [Agresti, 2002, Chap. 4]. Ce sont principalement la régression logistique, et dans une moindre mesure la régression linéaire multiple et la régression de Poisson, qui sont utilisées. Le modèle de Cox (appelé aussi modèle de survie), appliqué à des mesures temporelles de la variable à expliquer, n'est pas détaillé ici. La fréquence d'utilisation des différents modèles est donnée dans le tableau 2.1. Un inconvénient souvent évoqué est la sensibilité des modèles linéaires généralisés aux variables explicatives

corrélées entre elles, mais aucun article ne fait référence à des méthodes alternatives de type régression *PLS* linéaire généralisée pour pallier ce problème [Marx, 1996; Bastien *et al.*, 2005].

Nature de y (loi associée)	**Nature des X**	**Lien** (y, X)	Méthode	**Nb.citations** *PVM*
Dichotomique (binomiale)	Mixtes	*Logit*	Régr. logistique	249 articles (10.3%)
Ordinale (multinomiale)	Mixtes	*Logit* généralisée	Régr. log. multinomiale	
Dichotomique (binomiale)	Mixtes	*Log-Log*	Régr. Cox	108 articles (4.4%)
Quantitative (normale)	Quantitatives	Identité	Régr. linéaire mult.	57 articles (2.3%)
Effectif (Poisson)	Mixtes	*Log*	Régr. de Poisson	28 articles (1.1%)
Quantitative (normale)	Qualitatives	Identité	An. de variance	10 articles (0.4%)
–	Qualitatives	*Log-Log*	Modèle log-linéaire	5 articles (0.2%)
Quantitative (normale)	Mixtes	Identité	An. de covariance	4 articles (0.2%)

TAB. 2.1 – Utilisation des modèles linéaires généralisés dans les articles d'épidémiologie animale parus de la revue *Preventive Veterinary Medicine* (*PVM*).

La régression logistique [Hosmer et Lemeshow, 1989; Allison, 1999; Agresti, 2002] modélise une variable dichotomique ($y = 1$ indique la présence de la maladie et $y = 0$ son absence), ce qui permet un calcul direct de l'*odds ratio* ou du risque relatif, dont l'interprétation est précieuse en épidémiologie (paragraphe 1.2.2 page 25). Ceci justifie la mise en classes fréquente des variables quantitatives. Son utilisation, bien qu'elle soit souvent réductrice, est assez consensuelle parmi les épidémiologistes. Pour le cas où la variable à expliquer est ordinale ($y = 0$ caractérise les témoins, $y = 1$ les cas intermédiaires et $y = 2$ les cas sévères par exemple), la régression multinomiale est utilisée. Pour plus de détails sur l'utilisation de ces deux méthodes pour le traitement de données d'épidémiologie animale, le lecteur intéressé peut se référer à Dohoo *et al.* [2003, Chap. 16 et 17]. La régression linéaire multiple est utilisée dans les cas, assez rares, où la variable à expliquer est quantitative et n'a pas été recodée. Des exemples de traitements de données d'épidémiologie animale par cette méthode sont donnés par Dohoo *et al.* [2003, Chap. 14]. Les objectifs de la régression de Poisson sont d'estimer les effets de facteurs de risque sur des taux d'incidence ou de mortalité, la variable à expliquer étant dans ce cas une variable de comptage [Droesbeke *et al.*, 2005]. Dohoo *et al.* [2003, Chap. 18] donnent des exemples d'utilisation de cette méthode pour des données d'épidémiologie animale.

Généralisation par les modèles linéaires généralisés mixtes

Les différentes méthodes de régression précitées ne permettent pas de tenir compte de la structure complexe des observations. Elles sont basées sur l'hypothèse que les observations sont indépendantes les unes des autres, ce qui n'est pas toujours le cas pour les données d'épidémiologie animale. Comme détaillé dans le paragraphe 1.3.3 page 28, les animaux étudiés sont emboîtés dans leur portée et/ou dans la case où ils sont élevés, dans leur bande ou leur lot, puis dans leur élevage, et éventuellement dans leur région. Il est connu des épidémiologistes que d'un élevage à un autre, les résultats peuvent être différents. Or la conclusion qui est donnée en termes de facteurs de risque, doit prendre en compte ces différences et conclure au-delà de celles-ci. Il est donc intéressant que les variables structurelles (région,

élevage, bande, ...) puissent avoir des effets aléatoires sur la variable à expliquer. Des exemples de différents types de données d'épidémiologie animale hiérarchisées sont donnés par Dohoo *et al.* [2003, Chap. 20]. Les modèles linéaires généralisés mixtes, appelés aussi modèles multiniveaux ou modèles hiérarchisés, répondent à cette problématique [Agresti, 2002, chap. 12]. Ils permettent d'étendre les modèles linéaires généralisés à l'utilisation conjointe de variables explicatives considérées comme ayant un effet fixe ou aléatoire sur la variable à expliquer [Dohoo *et al.*, 2003, Chap. 21 et 22]. Au regard de la complexité des modèles ainsi définis, les modèles mixtes bayésiens sont de plus en plus utilisés pour pallier les limites des algorithmes d'estimation basés sur la vraisemblance [Dohoo *et al.*, 2003, Chap. 23]. Davantage de détails sur l'utilisation des modèles bayésiens dans ce contexte sont donnés par Congdon [2003, 2005].

Prise en compte de plusieurs enquêtes analytiques

D'autres voies sont explorées pour déterminer les facteurs de risque d'une maladie, en prenant en compte des données issues de plusieurs enquêtes analytiques d'épidémiologie animale. La première voie s'appuie sur une utilisation classique des statistiques bayésiennes : les enquêtes précédentes servent d'*a priori* pour quantifier le lien entre la variable à expliquer et les variables explicatives d'une nouvelle enquête par un modèle bayésien. La seconde voie permet de prendre en compte un ensemble d'enquêtes comparables, afin de déterminer les facteurs de risque de la maladie étudiée ; elle est vue au travers des méta-analyses [Dohoo *et al.*, 2003, Chap. 24].

Bilan sur l'utilisation de la régression en épidémiologie animale

Les techniques de régression appliquées au traitement des données d'épidémiologie animale analytique sont très fréquemment utilisées du fait de leurs nombreux avantages. Elles permettent tout d'abord de quantifier le lien entre une variable à expliquer (la maladie) et un ensemble de variables, potentiellement facteurs de risque à effet délétère ou protecteur. La quantification de ce lien, au travers de l'*odds ratio* ou du risque relatif, est une finalité en épidémiologie analytique. De plus, les modèles linéaires généralisés permettent de prendre en compte diverses natures de variables (tableau 2.1, page 32). L'extension de ces modèles aux modèles linéaires généralisés mixtes permet de prendre en compte des structures emboîtées d'individus, problématique classique en épidémiologie animale. Cependant, certains problèmes subsistent à l'utilisation de ces méthodes. Le problème le plus crucial est dû au grand nombre de variables X, multicorrélées, collectées lors des enquêtes. D'autres problèmes, liés à la caractérisation de la variable à expliquer par plusieurs variables, ou la prise en compte de l'organisation du questionnaire en groupes de variables de tailles déséquilibrées, subsistent. La résolution de ces problèmes peut passer avantageusement par l'utilisation de l'analyse de données, comme cela est détaillé dans le paragraphe suivant.

2.1.2 Recours à l'analyse de données

Le nombre d'articles relatifs à l'analyse de données dans le journal *Preventive Veterinary Medicine* est de 24, ce qui représente à peine 1% des articles parus. L'utilisation de l'analyse de données étant populaire en France, nous avons également fait des prospections dans la revue *Epidémiologie et Santé Animale*, qui est la revue de référence pour l'épidémiologie vétérinaire française. Seuls quatre articles (1%) ont eu recours à des méthodes d'analyse de données (sur un nombre estimé d'articles de l'ordre de 400 depuis 1982). Les différentes utilisations de l'analyse des données pour le traitement d'enquêtes d'épidémiologie animale sont détaillées par thème dans les paragraphes suivants.

Analyse de données à caractère descriptif

L'utilisation la plus classique de l'analyse des données est descriptive. Cette description peut aussi servir à réduire le nombre de variables X si celui-ci est trop important pour permettre la modélisation. L'article de Madec et Josse [1984] pose les fondements théoriques de l'utilisation descriptive de l'analyse des correspondances multiples, ou *ACM*, [Benzecri, 1973; Lebart *et al.*, 2000], pour décrire les liens au sein d'un ensemble de variables qualitatives, ainsi que le mode d'application au traitement des enquêtes épidémiologiques analytiques. Selon la nature des données, l'*ACM*, l'analyse en composantes principales, ou *ACP*, [Jolliffe, 1986; Lebart *et al.*, 2000], ou l'*ACP varimax* [Kaiser, 1958], sont utilisées pour décrire et comprendre les liens complexes entre les variables explicatives. Boklund *et al.* [2004] utilisent la procédure *prinqual* du logiciel *SAS* [SAS, 2004] pour intégrer directement les variables qualitatives et quantitatives à l'*ACP*. Les articles d'application utilisant l'analyse de données à caractère descriptif sont résumés dans le tableau 2.2.

Analyse de données en vue de recoder un ensemble de variables

Les méthodes d'analyse de données peuvent être utilisées pour synthétiser un ensemble de variables (tableau 2.3). Dans leur application classique, les méthodes usuelles de régression ne permettent pas la prise en compte de plusieurs variables à expliquer. Dans ce cas, il est intéressant d'en faire la synthèse en une seule variable. Dans les trois articles utilisant l'analyse de données avec cet objectif, une classification mixte sur composantes d'*ACM* des individus est réalisée [Lebart *et al.*, 2000]. Ce sont les appartenances aux classes qui servent de nouvelle variable à expliquer [Madec *et al.*, 1998; Rose *et al.*, 2003b], ou de nouvelle variable explicative [Aumont *et al.*, 1992], pour la suite des traitements. On peut noter que dans le cadre de la régression multivariée, le vecteur y peut être remplacé par le tableau Y, ce qui revient à réaliser une série de régressions simples.

Analyse de données en vue d'une régression

Les variables explicatives étant collectées en grand nombre lors des enquêtes épidémiologiques, le problème de leur multicolinéarité se pose lors de l'étape explicative. Une solution classique issue de l'analyse factorielle est d'utiliser les compo-

Référence	Objectif statistique	Objectif epidémiologique	Méthode
Madec et Josse [1984]	Descriptif	Pas d'application aux données épidémiologiques (article théorique)	*ACM*
Aumont *et al.* [1992]	Descriptif	Description du mode d'élevage de vaches, associé à des variables environnementales	*ACM*
Goodger *et al.* [1993]	Sélection de var.	Description des pratiques de conduite en élevage laitier	*ACP*
Hurnik *et al.* [1994]	Descriptif	Description de variables environnementales (en vue du lien avec la pathologie respiratoire du porc)	*ACP varimax*
O'Brien *et al.* [1995]	Descriptif	Description de la concentration en éléments chimiques dans le foie d'oiseaux aquatiques	*ACP*
Faye *et al.* [1997]	1.Descriptif 2.Multicolinéarité	Description des problèmes de santé des vaches laitières Variables relatives à la conduite d'élevage des vaches	*AFC* *ACM*
Duchateau *et al.* [1997]	Multicolinéarité (Sélection de var.)	Description de variables environnementales (en vue du lien avec l'infection de bovins au *Theileriosis*)	*ACP varimax*
Martrenchar *et al.* [2002]	Multicolinéarité (Sélection de var.)	Variables relatives à la conduite d'élevage de volailles (en vue du lien avec la prévalence de pododermatites)	*ACM*
Manske *et al.* [2002]	Grouper var.	Description des lésions et boiteries des vaches laitières	*ACP varimax*
Wagner *et al.* [2003]	Descriptif	Résistances conjointes d'E.Coli (issue de bovins) à différents antibiotiques	*ACP varimax*
Boklund *et al.* [2004]	Multicolinéarité (Sélection de var.)	Mise en évidence de grands thèmes dans les mesures de biosécurité au sein des élevages porcins	*ACP varimax*
Berghaus *et al.* [2005]	Multicolinéarité (Sélection de var.)	Sélection des principales variables liées aux pratiques d'élevage (pour le lien avec la paratuberculose bovine)	*ACP varimax*

TAB. 2.2 – Articles d'épidémiologie animale utilisant l'analyse de données à caractère descriptif.

Référence	Objectif statistique	Objectif epidémiologique	Méthode
Aumont *et al.* [1992]	Synthétiser plusieurs X	Définir une variable synthétisant la conduite d'élevage en vue d'expliquer l'infection de vaches aux *trichostrongyles*	Classif. sur comp. d'*ACM*
Madec *et al.* [1998]	Synthétiser plusieurs Y	Définir une variable caractérisant la sévérité des diarrhées du porcelet au sevrage	Classif. sur comp. d'*ACM*
Rose *et al.* [2003b]	Synthétiser plusieurs Y	Définir une variable caractérisant la qualité de désinfection des élevages de poulets de chair	Classif. sur comp. d'*ACM*

TAB. 2.3 – Articles d'épidémiologie animale utilisant l'analyse de données en vue de recoder un ensemble de variables.

santes qui en sont issues en tant que variables dans la régression. C'est le principe de la régression orthogonalisée [Massy, 1965]. L'article de Lafi et Kaneene [1992] pose les fondements théoriques de la régression sur composantes d'*ACP* dans le

cadre de l'épidémiologie animale. Selon la nature des variables X, la détermination des composantes est effectuée grâce à une *ACM*, une *ACP* ou une analyse factorielle permettant de traiter des variables quantitatives et qualitatives (procédure *prinqual* du logiciel *SAS* [SAS, 2004]). Dans certains modèles, des variables présentant un intérêt particulier, peuvent être associées aux composantes [Ganaba *et al.*, 1995; Ott *et al.*, 2003; Woods *et al.*, 2003]. Selon la nature de la variable à expliquer, le modèle est soit une régression linéaire multiple, soit une régression logistique [Hosmer et Lemeshow, 1989], soit une régression *tobit* [Tobin, 1958], soit un *path model* [Wright, 1921] pour les cas plus complexes. Dans la plupart des articles cités dans le tableau 2.4, les composantes sont interprétées comme des variables de synthèse. Les résultats sont généralement utilisés dans une perspective descriptive et rarement prédictive.

Référence	**Objectif statistique**	**Objectif epidémiologique**	**Méthode**
Lafi et Kaneene [1992]	Multicolinéarité	Pas d'application aux données épidémiologiques (article théorique)	Régr. sur comp. d'*ACP* (*PCR*)
Ganaba *et al.* [1995]	Multicolinéarité	Facteurs de risque de la mortalité et de la croissance des veaux	Path model sur var. et comp. d'*ACP*
Dohoo *et al.* [1997]	Multicolinéarité	Facteurs de risque de la pneumonie des porcs	Régr.logistique sur comp. d'*ACP*
Thoefner *et al.* [2001]	Multicolinéarité	Facteurs de risque de la colique du cheval	Régr.logistique sur comp. d'*ACP varimax*
Chi *et al.* [2002]	Multicolinéarité	Facteurs de risque de 4 maladies des bovins causant des pertes économiques	Régr. tobit sur comp. d'*ACP*
Ott *et al.* [2003]	Multicolinéarité	Lien entre la production laitière et la séro-prévalence liée l'infection à la leucose bovine, ainsi que les variables mesurant la conduite d'élevage	Régr. sur var. et comp. d'*ACM*
Woods *et al.* [2003]	Multicolinéarité	Lien entre les actions du fermier et les services et conseils du technicien vétérinaire	Path model sur var. et comp. d'*ACP varimax*
Berghaus *et al.* [2005]	Multicolinéarité	Facteurs de risque de la paratuberculose dans les élevages bovins laitiers	Régr. logistique sur comp. d'*ACP varimax*

Tab. 2.4 – Articles d'épidémiologie animale utilisant l'analyse de données en vue d'une régression.

Analyse de données à caractère explicatif

Une utilisation intéressante et appropriée de l'analyse des données pour le traitement des enquêtes d'épidémiologie animale est de permettre la compréhension des liens complexes entre les nombreuses variables explicatives et la (ou les) variable(s) à expliquer (tableau 2.5). Une utilisation simple et directe de l'analyse factorielle dans ce cas, consiste à effectuer une *ACM* sur les variables explicatives. Par la suite, la variable à expliquer est projetée en tant que variable supplémentaire [Faye et Lescourret, 1989; Ducrot et Cimarosti, 1991; Dohoo *et al.*, 1997; Madec *et al.*, 1998;

Thomsen *et al.*, 2007]. Cependant, cette technique a ses limites car les composantes de l'*ACM* ne sont pas orientées vers l'explication de Y, ce qui rend l'interprétation délicate. A ce titre, Dohoo *et al.* [1997] signalent que l'*ACM* ne permet pas de quantifier l'effet de facteurs de risque ; elle doit donc être utilisée en complément de la régression logistique ou d'autres méthodes de régression appropriées. Cette technique est présentée dans les articles de Dohoo *et al.* [1997]; Madec *et al.* [1998]; Thomsen *et al.* [2007]. L'utilisation de l'analyse factorielle discriminante, ou *AFD* [Fisher, 1936; Saporta, 2006; Celeux et Nakache, 1994], permet d'orienter les composantes vers l'explication d'une variable à expliquer qualitative. Dans ce cas, la contribution des variables explicatives à la construction des composantes permet une interprétation directe de celles-ci en tant que facteurs de risque de la maladie, ce qui est réalisé par Barnouin *et al.* [1995]. L'article de Morignat *et al.* [2006] illustre une utilisation de l'analyse discriminante à des fins explicatives mais aussi prédictives. Du fait de la nature particulière des variables à expliquer (effectifs), Lescourret et Faye [1991]; Faye *et al.* [1997] utilisent l'analyse canonique des correspondances, ou *ACC* [Ter Braak, 1986] pour quantifier la liaison entre les variables. Les techniques de segmentation [Celeux et Nakache, 1994] sont parfois utilisées. Elles ne sont pas évoquées ici car elles ne relèvent pas d'une analyse factorielle, qui est le concept qui a été privilégié dans le cadre de ce travail de recherche.

Référence	**Objectif statistique**	**Objectif epidémiologique**	**Méthode**
Faye et Lescourret [1989]	Décrire la liaison entre X et Y	Facteurs de risque des boiteries des vaches laitières	*ACM* avec Y sup.
Ducrot et Cimarosti [1991]	Décrire la liaison entre X et Y	Facteurs de risque de l'*Ecthyma* des ovins	*ACM* avec Y sup.
Lescourret et Faye [1991]	Quantifier la liaison entre X et Y	Facteurs de risque de la contamination du lait de vache	An. canonique des correspondances
Barnouin *et al.* [1995]	Quantifier la liaison entre X et Y	Facteurs de risque des mammites des vaches laitières	An. discriminante barycentrique
Dohoo *et al.* [1997]	Décrire la liaison entre X et Y	Facteurs de risque de la pneumonie des porcs	*ACM* avec Y sup.
Faye *et al.* [1997]	Quantifier la liaison entre X et Y	Facteurs de risque de différents problèmes de santé des vaches laitières	An. canonique des correspondances
Madec *et al.* [1998]	Décrire la liaison entre X et Y	Facteurs de risque des diarrhées du porcelet au sevrage	*ACM* avec Y sup.
Morignat *et al.* [2006]	Utiliser X pour prédire Y	Discrimination de 3 types de prions relatifs à l'*ESB* chez les bovins	An. discriminante
Thomsen *et al.* [2007]	Sélection de X orientée vers Y	Facteurs de risque d'un score élevé de réforme chez les vaches laitières	*ACM* avec Y sup.

TAB. 2.5 – Articles d'épidémiologie animale utilisant l'analyse de données à caractère explicatif.

Analyse de données multibloc

Une seule publication du journal *Preventive Veterinary Medicine* fait usage d'une méthode multibloc. Rougoor *et al.* [1999] utilisent l'approche *PLS* [Wold, 1982] pour expliquer la production laitière (respectivement la marge brute issue de cette production) par 31 variables explicatives (respectivement 22) réparties en 12 groupes (respectivement 10). Chaque groupe de variables explicatives est résumé par une variable latente ; les variables latentes sont soient liées entre elles, soient orientées vers l'explication de chacune des deux variables à expliquer. L'utilisation de l'approche *PLS* n'est pas justifiée par la structure multibloc des variables explicatives, mais par le nombre restreint d'individus comparé au nombre de variables explicatives.

Bilan sur l'utilisation de l'analyse des données en épidémiologie animale

L'analyse de données est, dans quasiment tous les cas, exceptés cinq cas réellement descriptifs, utilisée pour pallier les limites des techniques de régression. Elle sert à sélectionner les variables explicatives influentes en vue d'une modélisation, à synthétiser un ensemble de variables en une seule afin de permettre l'utilisation ultérieure de la régression, et enfin à résoudre le problème de la multicolinéarité des variables explicatives par les techniques de régression orthogonalisée. Dans certains rares cas, les techniques de régression sont même remplacées par des techniques d'analyse de données à caractère prédictif, telles que l'analyse discriminante ou l'analyse canonique des correspondances.

2.2 Problématique statistique en épidémiologie animale

2.2.1 Problèmes liés au grand nombre de variables explicatives

La multicolinéarité entre les variables explicatives est souvent rencontrée. Après un premier tri des variables explicatives les plus corrélées à la variable à expliquer, un second tri est réalisé sur les variables explicatives corrélées entre elles. Cependant, il est bien connu que des variables peuvent être faiblement corrélées deux à deux et présenter une multicolinéarité impliquant plusieurs variables [Weissfeld et Sereika, 1991; Dohoo *et al.*, 1997; Allison, 1999]. L'étude des corrélations entre les variables explicatives deux à deux est donc inadaptée pour appréhender les problèmes de corrélation complexes impliquant plus de deux variables, problématique courante en épidémiologie animale. Ces tris sont pourtant nécessaires, mais il faut savoir qu'ils réduisent mais ne suppriment pas la multicolinéarité. Une solution alternative peut être de créer de nouvelles variables, combinaison de plusieurs variables corrélées [Dohoo *et al.*, 1997]. Un exemple est donné sur un jeu de données d'épidémiologie animale. Les corrélations les plus significatives entre les 40 variables explicatives d'un jeu de données relatif à l'entérocolite épizootique du lapin (*EEL*), évoqué dans le paragraphe 1.3.2, page 26, sont illustrées par la figure 2.1. Parmi ces corrélations, 12.7% (soit 99/780 paires) sont significatives à 5% et concernent 38 variables sur 40. 4.5% sont significatives à 1% et concernent 28 variables sur 40. Sur la base de ces corrélations et de critères biologiques pertinents, 16 variables sont retenues. Cette

sélection réduit le nombre de paires corrélées au seuil de 5%. Le nombre de variables concernées par ces corrélations significatives reste grand, à savoir 13 variables sur 16.

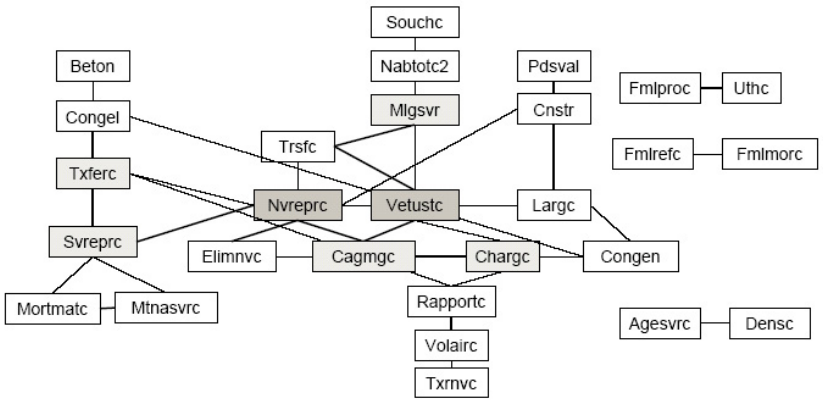


FIG. 2.1 – Illustration des corrélations entre les variables explicatives pour les données de l'enquête relative à l'*EEL* du lapin. Les traits entre les variables représentent les corrélations significatives à moins de 1% ; les traits plus épais pour celles à moins 0.1%. Les variables grisées sont celles qui sont les plus liées aux autres.

Le tableau 1.2 page 28 montre qu'habituellement, le nombre d'individus est plus grand que le nombre de variables explicatives. Cependant, pour permettre le calcul d'*odds ratio* ayant du sens et faciliter l'interprétation des liens avec la variable à expliquer, l'ensemble des variables explicatives est fréquemment mis en classe. Le nombre de degrés de liberté associé au modèle de régression augmente. En effet, si une variable qualitative comprend k modalités, elle utilise $(k-1)$ degrés de liberté dans la régression au lieu de n'en utiliser qu'un si elle restait quantitative. De ce fait, les variables sélectionnées lors des tris préalables ne peuvent donc pas être toutes prises en compte dans l'étape de régression.

2.2.2 Problèmes liés à la structure en groupe des variables explicatives

Les principaux groupes de variables explicatives ayant une signification zootechnique (pratiques d'hygiène, conduite d'élevage, ...) sont détaillées dans le paragraphe 1.3.1, page 26. Cependant, cette structure forte des variables explicatives au niveau biologique n'est pas valorisée actuellement dans les traitements statistiques usuels. Du fait du sens biologique de ces blocs, il apparaît important de mesurer à la fois l'importance des variables mais aussi des blocs de variables dans l'explication de la maladie.

Le nombre de variables collectées par groupe, souvent différent d'un groupe à un autre, pourrait influencer les facteurs de risque sélectionnés par la régression. En effet, une étude de simulation [Derksen et Keselman, 1992] dans laquelle les auteurs ont fait varier le degré de corrélation entre les variables X, le nombre de variables X, le pourcentage de variables influentes et de variables associées au bruit du modèle (*i.e.* variables non liées à la variable à expliquer), ainsi que la taille de l'échantillon,

montre que : [1] Le degré de multicolinéarité entre les variables X est le facteur qui joue le plus sur le nombre de variables influentes sélectionnées. En effet, plus la multicolinéarité augmente, plus le nombre de variables influentes sélectionnées par le modèle est petit, [2] Le nombre de variables X est le facteur le plus important pour la sélection du nombre de variables liées au bruit ; plus il augmente, plus le nombre de variables liées au bruit sélectionnées par le modèle augmente, [3] Le nombre d'individus a un effet positif sur le nombre de variables influentes sélectionnées, mais cet effet est étonnamment faible. Pour avoir un ordre d'idée, la situation la plus favorable (pas de multicolinéarité, peu de variables X, grand nombre d'individus) sélectionne 20% de variables qui sont en réalité du bruit. Dans la situation la plus défavorable (très forte multicolinéarité, grand nombre de variables et faible nombre d'individus), 74% de variables sélectionnées relèvent du bruit. Nous pouvons donc émettre l'hypothèse que plus le nombre de questions relatives à un groupe est grand, plus le nombre de facteurs de risque issus de ce groupe a des chances d'être élevé. Or, le nombre de variables d'un groupe n'est pas nécessairement le reflet de son importance comparativement aux autres groupes.

2.2.3 Problèmes liés à l'explication de plusieurs variables

Dans le cas où l'on dispose de plusieurs variables à expliquer, une première solution préconisée consiste à établir plusieurs modèles, liant chacune des variables à expliquer aux variables explicatives. Cette solution permet de déterminer les facteurs de risque relativement à chaque aspect complémentaire de la maladie. Cependant, il arrive que les facteurs de risque ne soient pas concordants, car les variables à expliquer illustrent des aspects différents, pas nécessairement corrélés, de la maladie. La synthèse des résultats devient délicate. Un exemple est donné sur le jeu de données relatif à la maladie de l'amaigrissement du porcelet. Les trois variables à expliquer décrivent la proportion d'animaux de l'élevage ayant séroconverti après une infection par le circovirus *PCV*2 : truies, porcelets et porcs à l'engrais. Les variables à expliquer relatives aux proportions de truies et de porcelets séropositifs sont liées (corrélation significative) et non corrélées à la proportion de porcs séropositifs. En effet, les truies et les porcelets sont élevés ensemble en maternité (au préalable de la prise de sang qui a lieu lorsque les porcelets sont en post-sevrage, séparés de leur mère), ce qui explique que leurs proportions de séroprévalence soient comparables. Les porcs à l'engrais, plus âgés, ont perdu trace de l'immunité colostrale reçue de la truie. Ils sont élevés dans d'autres bâtiments et sont donc soumis à un milieu différent. Pour déterminer les facteurs de risque de la maladie au niveau de l'élevage, il est essentiel de raisonner sur l'ensemble des variables à expliquer. L'objectif à atteindre est un élevage où la proportion de truies et de porcelets ayant séroconverti est élevée (transmission d'anticorps de la mère au porcelet par le colostrum), et où la proportion de porcs à l'engrais ayant séroconverti est faible (pas de contact avec le circovirus *PCV*2). Ce qui revient à dire qu'un facteur de risque est ici un facteur diminuant la proportion de truies et de porcelets et augmentant la proportion de porcs à l'engrais ayant séroconverti [Rose *et al.*, 2003a]. Nous constatons ici que le fait de raisonner sur chacune des variables à expliquer ne permet pas de répondre correctement à la question posée

par l'épidémiologiste.

Une deuxième solution consiste à créer une variable de synthèse. Dans ce cas, la maladie ou le phénomène de santé étudié est caractérisé par la combinaison de plusieurs variables. Ces variables sont parfois synthétisées en une seule variable, en les croisant simplement ou par une analyse factorielle couplée à une classification sur les composantes (paragraphe 2.1.2, page 34). Ce sont les résultats de la classification des individus qui servent à la codification de la nouvelle variable à expliquer. Ces synthèses conduisent souvent à une trop grande simplification des données et donc à une perte d'information peu satisfaisante pour l'épidémiologiste. Un exemple est donné sur un jeu de données d'épidémiologie animale. L'entérocolite épizootique du lapin (*EEL*) est une maladie se traduisant par des symptômes (perte d'appétit, ballonnements, diarrhées, . . .), ainsi que par des lésions plus ou moins prononcées se détectant à l'autopsie de l'animal (estomac, cæcum, intestin principalement) [Klein, 2002]. Ces lésions peuvent conduire à la mort de l'animal. L'enquête analytique *EEL* évoquée dans le paragraphe 1.3.2 page 26, est basée sur quatre variables à expliquer pour déterminer si l'élevage est cas, témoin ou a un statut intermédiaire : le pourcentage de mortalité ainsi que les signes d'*EEL* à l'engraissement et en maternité lors des cinq dernières bandes d'animaux. Le codage de la variable à expliquer, synthétique, est détaillé dans la figure 2.2. La synthèse de cette information riche, réalisée par un groupe d'experts, reste empirique et évidemment réductrice.

<table>
<tr><th></th><th>CAS</th><th colspan="2">TEMOINS</th><th colspan="2">ELEVAGES INTERMEDIAIRES</th></tr>
<tr><td rowspan="3">Critères de sélection</td><td rowspan="2">Présence de signes d'EEL sur les 5 dernières bandes (5 « oui »)</td><td rowspan="2">Absence de signes d'EEL sur les 5 dernières bandes (5 « non »)
Mortalité < 10 % sur chaque en maternité et en engraissement</td><td rowspan="2">9</td><td>Présence de l'EEL inconnue en engraissement sur une ou plusieurs bandes</td><td>8</td></tr>
<tr><td>Présence et absence de l'EEL inconstantes d'une bande à l'autre</td><td>13</td></tr>
<tr><td>Mortalité en engraissement > 12 % en moyenne sur les 5 dernières bandes</td><td>Mortalité élevée en engraissement et / ou en maternité si et seulement si elle n'est pas due à l'EEL</td><td>19</td><td>Présence de signes d'EEL sur les 5 dernières bandes mais mortalité moyenne sur les 3 dernières bandes < 12 %</td><td>9</td></tr>
<tr><td>Nombre total d'élevages</td><td>37</td><td colspan="2">28</td><td colspan="2">Absence de signes d'EEL sur les 5 dernières bandes mais signes d'EEL en maternité
31</td><td>1</td></tr>
</table>

FIG. 2.2 – Définition d'une variable Y de synthèse pour l'enquête sur l'*EEL* du lapin, d'après Klein [2002].

2.3 Contexte du travail de recherche

Le développement de méthodes statistiques permettant de traiter les données d'épidémiologie animale doit prendre en compte la particularité de ces données, ainsi que les objectifs de traitement associés. Il est illusoire de résoudre à la fois tous les problèmes statistiques liés à leur traitement. Cinq contraintes sont explorées, répondant aux problématiques majeures de l'étape indispensable de modélisation (paragraphe 2.1.1 page 31) :

- Problématique [1] : prise en compte d'un grand nombre de variables explicatives, parfois supérieur au nombre d'individus, et liées entre elles,
- Problématique [2] : prise en compte de plusieurs variables à expliquer,
- Problématique [3] : prise en compte de blocs de variables explicatives, de tailles déséquilibrées, ayant un sens biologique dont l'interprétation est intéressante,
- Problématique [4] : modélisation du lien entre les variables explicatives et les variables à expliquer, en tenant compte des trois problématiques précédentes,
- Problématique [5] : les variables explicatives et à expliquer sont de nature mixte (qualitatives et quantitatives).

Le choix qui est fait pour répondre à ces problématiques est orienté vers les méthodes relevant de l'analyse factorielle. En effet, l'utilisation de méthodes factorielles répond tout d'abord à la problématique [1], à savoir la prise en compte d'un grand nombre de variables explicatives liées entre elles. L'utilisation de ces méthodes répond à l'objectif descriptif initial du traitement des données d'épidémiologie animale : comprendre le mode de fonctionnement de l'ensemble des élevages étudiés (par l'étude des liens entre les variables), et visualiser les ressemblances et différences entre les élevages (par l'étude des individus). De plus, l'analyse factorielle permet une description adaptée des données d'épidémiologie, car son objectif est de décrire un vaste ensemble de variables explicatives ayant potentiellement une influence, jouant toutes le même rôle, inter-agissant entre elles, dont on ne connaît *a priori* pas la structure. Pour cela, l'analyse factorielle décrit et synthétise l'information contenue dans le tableau de données. La sélection des variables explicatives, nécessaire et contraignante pour les techniques de régression usuelles, n'est plus nécessaire ; l'analyse factorielle permet donc de prendre en compte une information plus riche. Les résultats de l'analyse factorielle sont souvent données sous forme de représentations graphiques synthétiques qui permettent d'appréhender les liens complexes unissant un grand nombre de variables et/ou d'individus [Lebart *et al.*, 2000]. L'exploration des liens entre individus ainsi que les liens entre individus et variables, apporte une information intéressante à l'épidémiologiste.

L'utilisation de méthodes factorielles décrivant les variables explicatives est intéressante, mais n'est pas parfaitement adaptée aux objectifs de traitement des données d'épidémiologie animale analytique. En fait, c'est la description des liens entre les variables explicatives et des variables liées à la maladie, qui intéresse réellement l'épidémiologiste. Nous nous situons ici dans le champ des méthodes factorielles permettant la comparaison de deux tableaux, ayant ses origines dans les travaux relatifs à l'analyse canonique [Hotelling, 1936]. Cette méthode, au rôle théorique important, permet d'établir un pont entre les méthodes descriptives et les méthodes explicatives. Elle admet d'ailleurs pour cas particuliers des méthodes descriptives (l'*AFC* par exemple) et explicatives (la régression linéaire multiple, par exemple). Les principales méthodes, permettant de lier deux tableaux, qui en sont issues, à savoir l'analyse factorielle inter-batterie [Tucker, 1958], l'analyse en composantes principales sur variables instrumentales [Rao, 1964] et la régression *PLS* [Wold, 1966], sont détaillées dans la partie II page 45. Ces méthodes permettent de prendre en compte plusieurs variables Y, ce qui répond à la problématique [2].

La recherche a été très active ces dernières décennies dans le domaine de l'analyse factorielle de données structurées en plusieurs blocs. La prise en compte de variables explicatives structurées en plusieurs blocs est possible grâce au cadre théorique donné par la généralisation de l'analyse canonique [Horst, 1961; Carroll, 1968]. L'analyse canonique généralisée permet de représenter les variables de chaque bloc dans une base, commune à l'ensemble des variables explicatives, dont la liaison avec chacun des blocs est optimisée. Cette représentation factorielle synthétique permet de mesurer l'influence de chaque bloc dans la construction de l'espace commun. Chaque bloc a le même poids initial, ce qui permet de s'affranchir du nombre déséquilibré de variables au sein de chaque bloc. L'utilisation de ces tableaux multiples répond à une partie des objectifs formulés par la problématique [3]. Il est de plus possible d'orienter la description de K tableaux vers un $(K+1)^{ieme}$, en l'occurrence le tableau à expliquer décrivant la maladie. Des informations nouvelles sont ainsi apportées à l'épidémiologiste : il devient possible de mesurer l'influence de chaque bloc de variables explicatives dans l'explication de la maladie. Il est bien entendu toujours possible d'avoir des informations au niveau des variables. Ces méthodes $(K+1)-$tableaux répondent à la problématique [3]. Elles sont détaillées dans la partie III page 89.

La régression est l'étape [4] indispensable du traitement des données d'épidémiologie animale, qui permet d'expliquer et quantifier le lien entre la variable à expliquer et les variables explicatives. Les techniques de régression orthogonalisée [Massy, 1965; Saporta, 1975] associent régression et analyse factorielle, en remplaçant dans la régression les variables explicatives par les composantes issues de l'analyse factorielle. Le problème de la multicolinéarité est nettement diminué si les composantes considérées comme étant liées au bruit sont écartées de la régression [Lebart *et al.*, 2000; Barker et Brown, 2001]. Le nombre de composantes à conserver est un compromis entre la réduction de la variance des coefficients de régression et l'introduction d'un biais trop important. Le retour aux coefficients de régression des variables, à partir de ceux des composantes, est aisé. L'utilisation de la régression orthogonalisée est donc une solution qui permet de limiter le problème de la multicolinéarité des variables explicatives, et ainsi d'en prendre plus en compte pour expliquer la maladie. Ces deux avantages sont satisfaisants pour l'épidémiologiste, et répondent à la problématique [4].

La dernière problématique [5] concerne la prise en compte des variables de nature mixte. Des variables de nature mixte correspondent à un mélange de variables quantitatives et de variables qualitatives de différentes natures (dichotomique, nominale, ordinale), qui concernent les variables explicatives mais aussi les variables à expliquer. Dans ce travail de recherche, les variables qualitatives, dans la plupart des cas dichotomiques, font l'objet d'un codage disjonctif complet, en supprimant une des modalités [Lebart *et al.*, 2000, p. 238]. Elles sont ainsi associées au traitement des variables quantitatives. Cette pratique est usuelle ; se référer à Tenenhaus [1998, Chap. 12] à titre d'exemple pour la régression *PLS*.

Deuxième partie

Description d'un tableau X orientée vers l'explication d'un tableau Y

Chapitre 3

Analyse de deux tableaux

3.1 Méthodes liant deux tableaux X et Y

Nous disposons d'un tableau X constitué de P variables explicatives $[x_1, \ldots, x_P]$ et d'un tableau Y formé de Q variables à expliquer $[y_1, \ldots, y_Q]$. Ces variables sont mesurées sur les mêmes N individus. Par la suite, ces variables sont supposées centrées. La structure des données est illustrée par la figure 3.1.

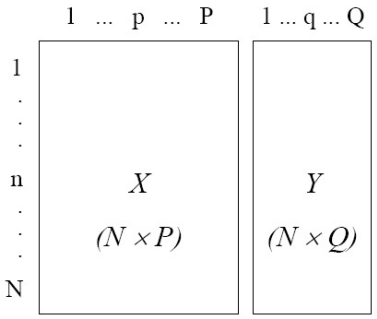


FIG. 3.1 – Illustration de la structure des tableaux X et Y.

3.1.1 Analyse en composantes principales sur variables instrumentales

L'objectif étant d'expliquer plusieurs variables Y à partir de variables X, la première méthode préconisée est une généralisation de la régression linéaire multiple. Cette généralisation peut être vue à travers l'analyse en composantes principales sur variables instrumentales, ou *ACPVI* [Rao, 1964]. La détermination de la première composante de l'*ACPVI* telle qu'elle est proposée par Rao [1964] est basée sur la minimisation du critère (3.1). Par la suite, les normes utilisées sont considérées comme étant des normes $L2$.

$$||YY' - \lambda^{(1)} t^{(1)} t^{(1)'}||^2 \quad \text{avec} \quad t^{(1)} = Xw^{(1)} \quad \text{et} \quad ||t^{(1)}|| = 1 \tag{3.1}$$

Concrètement, ceci revient à minimiser la distance entre la matrice de produits scalaires entre individus dans l'espace des variables de Y et la représentation des

individus sur la première composante $t^{(1)}$ contrainte d'être dans l'espace engendré par les variables de X. L'*ACPVI* a été étudiée de manière plus approfondie par Van Den Wollenberg [1977] et Sabatier [1987]. Ces auteurs ont en particulier développé plusieurs propriétés de cette méthode ainsi que des formulations montrant son positionnement par rapport à l'*ACP* et l'analyse canonique. Nous retenons ici la formulation qui fait apparaître l'*ACPVI* de Y par rapport à X comme une *ACP* de Y sous la contrainte que les composantes principales t sont des combinaisons linéaires des variables constituant X (variables instrumentales). En effet, la première composante $t^{(1)}$ de l'*ACPVI* de Y par rapport à X peut être déterminée de manière à maximiser le critère (3.2).

$$\sum_{q=1}^{Q} cov^2(y_q, t^{(1)}) \quad \text{avec} \quad t^{(1)} = Xw^{(1)} \quad \text{et} \quad ||t^{(1)}|| = 1 \tag{3.2}$$

Ce problème a pour solution $w^{(1)}$, vecteur propre de la matrice $[(1/N^2)(X'X)^{-1}X'YY'X]$ associé à la plus grande valeur propre $\lambda^{(1)}$. Les composantes de l'*ACPVI* d'ordre supérieur à un sont obtenues par une démarche itérative consistant, à chaque pas, à déterminer une composante t standardisée, maximisant le même critère, en imposant la contrainte supplémentaire que cette composante soit orthogonale aux composantes déterminées lors des étapes précédentes. La composante $t^{(h)} = Xw^{(h)}$ d'ordre h de l'*ACPVI* est donc associée au vecteur $w^{(h)}$, h^{ieme} vecteur propre de la matrice $[(1/N^2)(X'X)^{-1}X'YY'X]$, ce qui permet d'écrire que $(X'X)^{-1}X'YY'Xw^{(h)} = \lambda^{(h)}w^{(h)}$, ou de manière équivalente $X'YY'Xw^{(h)} = \lambda^{(h)}X'Xw^{(h)}$, pour $h = (1,\ldots,H)$. Par souci de cohérence avec les méthodes qui seront introduites par la suite, nous proposons de déterminer les composantes selon la procédure préconisée dans le cadre de la régression *PLS* (paragraphe 3.1.3 page 54), qui consiste à considérer, à chaque pas, les résidus de la régression des variables de X sur les composantes déterminées aux étapes précédentes. Ces deux démarches, issues soit des vecteurs propres successifs d'une même matrice ou soit de déflations, conduisent en réalité aux mêmes composantes. Pour cela, nous démontrons que chaque composante $t^{(h)}$ correspond à la première composante de l'*ACPVI* calculée sur $X^{(h-1)}$, résidu de régression de X à partir de $(t^{(1)},\ldots,t^{(h-1)})$. De manière plus précise, nous allons montrer que :

$$t^{(h)} = X^{(h-1)}w^{(h)} \tag{3.3}$$

$$X^{(h-1)'}YY'X^{(h-1)}w^{(h)} = \lambda^{(h)}X^{(h-1)'}X^{(h-1)}w^{(h)} \tag{3.4}$$

Comme nous savons que $X^{(h-1)} = \left[I - \sum_{j=1}^{h-1} \frac{t^{(j)}t^{(j)'}}{||t^{(j)}||^2}\right]X$, la preuve de l'égalité (3.3) est directe :

$$\begin{aligned} X^{(h-1)}w^{(h)} &= \left[Xw^{(h)} - \left(\sum_{j=1}^{h-1} \frac{t^{(j)}t^{(j)'}}{||t^{(j)}||^2}\right)Xw^{(h)}\right] \\ &= \left[t^{(h)} - \sum_{j=1}^{h-1} t^{(j)}\frac{t^{(j)'}t^{(h)}}{||t^{(j)}||^2}\right] \\ &= t^{(h)} \end{aligned}$$

La dernière égalité découle du fait que les composantes de l'*ACPVI* sont mutuellement orthogonales. Pour démontrer l'égalité (3.4), nous considérons que :

$$\begin{aligned}
X^{(h-1)'}YY'X^{(h-1)}w^{(h)} &= X'\left[I-\sum_{j=1}^{h-1}\frac{t^{(j)}t^{(j)'}}{||t^{(j)}||^2}\right]YY'X^{(h-1)}w^{(h)} \\
&= X'\left[I-\sum_{j=1}^{h-1}\frac{t^{(j)}t^{(j)'}}{||t^{(j)}||^2}\right]YY'Xw^{(h)} \\
&= X'YY'Xw^{(h)}-X'\left[\sum_{j=1}^{h-1}\frac{t^{(j)}}{||t^{(j)}||^2}\right]w^{(j)'}X'YY'Xw^{(h)} \\
&= \lambda^{(h)}X'Xw^{(h)}-X'\left[\sum_{j=1}^{h-1}\frac{t^{(j)}}{||t^{(j)}||^2}\right]w^{(j)'}\lambda^{(h)}X'Xw^{(h)} \\
&= \lambda^{(h)}X'Xw^{(h)}-\lambda^{(h)}X'\left[\sum_{j=1}^{h-1}\frac{t^{(j)}}{||t^{(j)}||^2}\right]t^{(j)'}t^{(h)} \\
&= \lambda^{(h)}X'Xw^{(h)} \qquad (3.5)
\end{aligned}$$

En ré-utilisant l'égalité précédente $X^{(h-1)}w^{(h)} = Xw^{(h)}$, nous avons :

$$\begin{aligned}
X^{(h-1)'}X^{(h-1)}w^{(h)} &= X^{(h-1)'}Xw^{(h)} \\
&= X'\left[I-\sum_{j=1}^{h-1}\frac{t^{(j)}t^{(j)'}}{||t^{(j)}||^2}\right]t^{(h)} \\
&= X'\left[t^{(h)}-\sum_{j=1}^{h-1}t^{(j)}\frac{t^{(j)'}t^{(h)}}{||t^{(j)}||^2}\right] \\
&= X't^{(h)} \\
&= X'Xw^{(h)} \qquad (3.6)
\end{aligned}$$

Des équations (3.5) et (3.6), il découle que :

$$X^{(h-1)'}YY'X^{(h-1)}w^{(h)} = \lambda^{(h)}X^{(h-1)'}X^{(h-1)}w^{(h)}$$

Ce qui démontre précisément l'égalité (3.4). Ceci permet de montrer que les composantes de l'*ACPVI* peuvent s'obtenir par déflation de la matrice X sur les composantes t précédemment obtenues.

Nous pouvons également montrer que les composantes t, solutions du problème d'optimisation (3.2), sont également solutions du problème (3.7) [Johansson, 1981], consistant à déterminer deux variables latentes, $t^{(1)}$ et $u^{(1)}$, de manière à maximiser le critère (3.7).

$$cov^2(u^{(1)},t^{(1)}) \quad \text{avec} \quad t^{(1)} = Xw^{(1)}, \quad u^{(1)} = Yv^{(1)}, \quad ||t^{(1)}|| = ||v^{(1)}|| = 1 \qquad (3.7)$$

Afin de démontrer cette propriété, nous remarquons que $cov^2(u^{(1)},t^{(1)}) = (1/N^2)(u^{(1)'}t^{(1)})^2 = (1/N^2)(v^{(1)'}Y't^{(1)})^2$. Le maximum, pour $v^{(1)}$, de cette quantité est atteint pour $v^{(1)} =$

$Y't^{(1)}/\|Y't^{(1)}\|$, ce qui permet d'obtenir en définitive $cov^2(u^{(1)}, t^{(1)}) = (1/N^2)(t^{(1)'}YY't^{(1)}) = \sum_q cov^2(y_q, t^{(1)})$. Nous reconnaissons dans cette dernière expression le critère (3.2). Ceci corrobore les liens déjà démontrés entre l'*ACPVI*, la régression *PLS* et l'analyse inter-batterie de Tucker [Van de Geer, 1984; Chessel et Mercier, 1993; Burnham *et al.*, 1996].

La méthode dite *reduced rank regression* [Muller, 1981; Davies et Tso, 1982] préconise que chaque variable du tableau Y soit modélisée par la régression sur les composantes $(t^{(1)}, \ldots, t^{(h)})$ déterminées par l'*ACPVI*. Il faut noter que comme la solution de l'*ACPVI* est basée sur l'inversion de la matrice $(X'X)$, des problèmes d'instabilité peuvent surgir en présence de quasi-colinéarité entre les variables du tableau X, comme cela est d'ailleurs le cas pour la régression linéaire multiple dont elle est une généralisation [Cazes, 1975]. Les méthodes présentées dans les paragraphes suivants 3.1.2 et 3.1.3 sont des méthodes présentant moins de sensibilité aux multicolinéarités entre les variables explicatives.

3.1.2 Méthodes issues de l'analyse en composantes principales

Régression sur composantes principales

La régression sur composantes principales utilise les composantes orthogonales de l'analyse en composantes principales pour pallier les problèmes de quasi-colinéarité entre les variables X. L'objectif de l'analyse en composantes principales, ou *ACP* [Jolliffe, 1986; Lebart *et al.*, 2000], est de décrire un tableau X contenant P variables $X = [x_1, \ldots, x_P]$. Pour cela, la méthode recherche des composantes t, combinaisons linéaires des variables X, décrivant au mieux les données. La solution de premier ordre de l'*ACP* consiste à maximiser le critère (3.8).

$$\begin{aligned} &var(t^{(1)}) \quad \text{avec} \quad t^{(1)} = Xw^{(1)} \quad \text{et} \quad \|w^{(1)}\| = 1 \\ &\qquad \text{ou} \\ &\sum_{p=1}^{P} cov^2(x_p, t^{(1)}) \quad \text{avec} \quad t^{(1)} = Xw^{(1)} \quad \text{et} \quad \|w^{(1)}\| = 1 \end{aligned} \tag{3.8}$$

La solution optimale est obtenue par $w^{(1)}$, vecteur propre de la matrice $(1/N)(X'X)$ associé à la plus grande valeur propre $\lambda^{(1)}$. La solution d'ordre deux, $w^{(2)}$, est le second vecteur propre de la même matrice, associé à la seconde valeur propre, et ainsi de suite.

Il est possible d'utiliser les composantes $(t^{(1)}, \ldots, t^{(h)})$, combinaisons linéaires des variables X, mutuellement orthogonales pour expliquer une ou plusieurs variables Y en réalisant une régression linéaire sur les composantes principales, appelée régression orthogonalisée, ou *PCR* [Massy, 1965; Tomassone *et al.*, 1983; Palm et Iemma, 1995]. L'inconvénient d'utiliser les composantes de l'*ACP* dans la régression est que celles-ci ne sont pas nécessairement orientées vers l'explication des variables Y. Il n'y a généralement pas de correspondance entre la grandeur de la valeur propre associée à une composante et son pouvoir explicatif [Vigneau *et al.*, 1996].

Latent root regression

Lorsque le tableau à prédire est unidimensionnel, soit $Y = [y]$, une solution permettant d'orienter les composantes t vers l'explication d'une variable y, est donnée par la *latent root regression*, ou *LRR* [Webster *et al.*, 1974]. Cette méthode est basée sur une *ACP* du tableau concaténé horizontalement $A = [y|X] = [y, x_1, \ldots, x_P]$. La composante $c^{(1)}$ de cette *ACP* correspond à la maximisation du critère (3.9).

$$cov^2(y, c^{(1)}) + \sum_{p=1}^{P} cov^2(x_p, c^{(1)}) \quad \text{avec} \quad c^{(1)} = Aa^{(1)} \quad \text{et} \quad \|a^{(1)}\| = 1 \tag{3.9}$$

La solution optimale est obtenue par $a^{(1)}$, vecteur propre de la matrice $(1/N)(A'A)$ associé à la plus grande valeur propre $\lambda^{(1)}$. Les solutions d'ordre suivant $(a^{(2)}, \ldots, a^{(h)})$ sont issues de la décomposition spectrale de la même matrice. A partir de chaque composante $c^{(h)} = a_0^{(h)} y + a_1^{(h)} x_1 + \cdots + a_P^{(h)} x_P$, sont définies les variables latentes $t^{(h)} = a_1^{(h)} x_1 + \cdots + a_P^{(h)} x_P$, combinaisons linéaires des variables X, utilisées pour la prédiction de la variable y. La stratégie adoptée par Webster *et al.* [1974] consiste à ne pas conserver toutes les composantes $(t^{(1)}, \ldots, t^{(h)})$ pour expliquer y. Le tri des composantes non prédictives est réalisé sur les valeurs propres $(\lambda^{(1)}, \ldots, \lambda^{(h)})$ associées aux composantes c, et sur les coefficients $(a_0^{(1)}, \ldots, a_0^{(h)})$ associés à la variable y. Les composantes ne sont pas conservées lorsque les valeurs propres $\lambda^{(h)}$ sont faibles ou que les coefficients $a_0^{(h)}$ ont des valeurs peu élevées [Palm et Iemma, 1995]. Cependant en pratique, ces critères se révèlent peu efficaces [Bertrand *et al.*, 2001].

Nous proposons une version modifiée de la *LRR*, permettant d'extraire directement les composantes t d'une décomposition spectrale. La stratégie consiste à changer le problème d'optimisation (3.9) en considérant le même critère, mais en imposant cette fois-ci aux composantes d'être une combinaison linéaire des variables X [Bougeard *et al.*, 2007]. Cette méthode peut être interprétée comme une *ACP* du tableau $A = [y|X]$ où les composantes sont contraintes d'être dans l'espace des variables X. Cette nouvelle formulation de la *LRR* permet d'extraire des composantes, orientées vers l'explication de y et directement utilisables pour l'expliquer. De plus, cette méthode peut être étendue au cas où l'on dispose de plusieurs variables Y, comme cela est détaillé dans le critère (3.10) qui consiste à maximiser :

$$\sum_{q=1}^{Q} cov^2(y_q, t^{(1)}) + \sum_{p=1}^{P} cov^2(x_p, t^{(1)}) \quad \text{avec} \quad t^{(1)} = Xw^{(1)}, \quad \|w^{(1)}\| = 1 \tag{3.10}$$

Le vecteur $w^{(1)}$ est le premier vecteur propre de la matrice $(1/N^2)(X'YY'X + X'XX'X)$.

En effet, le critère (3.10) peut s'écrire de la façon suivante :

$$\frac{1}{N^2}\left[\sum_q t^{(1)'} y_q y_q' t^{(1)} + \sum_p t^{(1)'} x_p x_p' t^{(1)}\right]$$

$$\frac{1}{N^2}\left[w^{(1)'} X' \sum_q (y_q y_q') X w^{(1)} + w^{(1)'} X' \sum_p (x_p x_p') X w^{(1)}\right]$$

$$\frac{1}{N^2}\left[w^{(1)'} (X'YY'X + X'XX'X) w^{(1)}\right]$$

Afin de déterminer la composante d'ordre deux, les tableaux X et Y sont remplacés dans le critère (3.10) par leurs résidus respectifs de la régression sur la première composante $t^{(1)} : X^{(1)} = \left(I - (t^{(1)} t^{(1)'})/(t^{(1)'} t^{(1)})\right) X$ et $Y^{(1)} = \left(I - (t^{(1)} t^{(1)'})/(t^{(1)'} t^{(1)})\right) Y$. Cette procédure est répétée plusieurs fois pour obtenir les composantes $(t^{(1)}, \ldots, t^{(h)})$. Cette procédure de déflation, utilisée classiquement pour la régression *PLS* [Wold *et al.*, 1983; Burnham *et al.*, 1996], conduit à l'obtention de composantes orthogonales mutuellement par construction, qui, de proche en proche, restituent la variabilité du tableau Y. Les composantes $(t^{(1)}, \ldots, t^{(h)})$ ainsi obtenues peuvent servir à des fins de prédiction, en régressant les variables Y sur celles-ci. Les avantages de cette version modifiée de la *LRR* sont que : 4 (paragraphe 6.1.5 page 104).

3.1.3 De l'analyse canonique à la régression *PLS*

Cadre théorique de l'analyse canonique

A titre de comparaison avec les autres méthodes, l'analyse canonique est présentée. Cependant, le fait qu'elle attribue un rôle symétrique aux tableaux X et Y et sa forte sensibilité à la multicolinéarité des variables la rend peu utilisable dans le cadre du traitement des données d'épidémiologie animale. L'analyse canonique proposée par Hotelling [1936] est une méthode qui permet de synthétiser les relations entre deux groupes de variables X et Y. Pour cela, la méthode recherche les combinaisons linéaires des variables X qui sont successivement les plus corrélées à des combinaisons linéaires des variables Y. De manière plus précise, la première étape de l'analyse canonique consiste à maximiser le critère (3.11).

$$cov^2(t^{(1)}, u^{(1)}) \quad \text{avec} \quad t^{(1)} = Xw^{(1)}, \quad u^{(1)} = Yv^{(1)}, \quad ||t^{(1)}|| = ||u^{(1)}|| = 1 \tag{3.11}$$

La solution de cette maximisation est donnée par $w^{(1)}$, vecteur propre de la matrice $(1/N^2)(X'X)^{-1}X'Y(Y'Y)^{-1}Y'X$ associé à la plus grande valeur propre $\lambda^{(1)}$, et $v^{(1)}$, premier vecteur propre de la matrice $(1/N^2)(Y'Y)^{-1}Y'X(X'X)^{-1}X'Y$, associé à la même valeur propre [Van Den Wollenberg, 1977]. Les solutions d'ordre suivant sont obtenues à partir des vecteurs propres successifs des mêmes matrices.

Du point de vue géométrique, l'analyse canonique revient à minimiser, dimension par dimension, l'angle entre la composante t, située dans l'espace des variables X, et la composante u, située dans l'espace des variables Y [Cazes, 1980]. Il faut noter que la recherche des couples de variables t et u de corrélation maximale, donne des composantes canoniques bien corrélées entre elles, mais souvent peu explicatives de leur groupe d'origine, l'inertie de chacun des groupes n'étant pas prise en

compte par le critère (3.11) [Gleason, 1976; Tenenhaus, 1998]. De plus, des corrélations canoniques peuvent apparaître artificiellement élevées du fait d'une variable x et d'une variable y très corrélées [Obadia, 1978]. Il est donc risqué de se baser sur les corrélations canoniques pour juger globalement les liaisons entre les tableaux X et Y.

Dans le cas d'une seule variable y à expliquer, l'analyse canonique revient à une régression multiple [Obadia, 1978] ; pour une explication détaillée, se référer à Lebart *et al.* [2000, p. 229]. La régression des corrélations canoniques proposée par Burnham *et al.* [1996] modélise les variables Y à partir des premières composantes $t = Xw$. Cependant, les premières composantes canoniques de X ne sont pas nécessairement les combinaisons linéaires les plus explicatives des variations des variables Y [Obadia, 1978]. Cette méthode est à la fois sensible à la quasi-colinéarité entre les variables X (inversion de la matrice $X'X$) et entre les variables Y (inversion de la matrice $Y'Y$). L'interprétation des résultats de cette méthode risque d'être délicate. Elle est peu usitée dans la pratique, mais son cadre théorique est fondamental.

Analyse factorielle inter-batterie et analyse de concordance

Tucker [1958] propose une modification du critère (3.11) de l'analyse canonique en basant le critère à maximiser sur une covariance et non plus sur une corrélation. Les composantes t et u recherchées sont à la fois corrélées entre elles, mais aussi représentatives de leur groupe de variables, du fait de la prise en compte de l'inertie de chacun des deux tableaux. L'analyse factorielle inter-batterie peut donc être vue comme un compromis entre l'analyse canonique des tableaux X et Y, l'*ACP* de X et l'*ACP* de Y [Frank et Friedman, 1993; Barker et Rayens, 2003]. Le critère à maximiser (3.12) de l'analyse factorielle inter-batterie devient :

$$cov(t^{(1)}, u^{(1)}) \quad \text{avec} \quad t^{(1)} = Xw^{(1)}, \quad u^{(1)} = Yv^{(1)}, \quad \|w^{(1)}\| = \|v^{(1)}\| = 1 \tag{3.12}$$

La solution de cette maximisation est donnée par $w^{(1)}$ premier vecteur propre de la matrice $(1/N)X'YY'X$ associé à la plus grande valeur propre $\lambda^{(1)}$, et $v^{(1)}$ premier vecteur propre normé de la matrice $(1/N)Y'XX'Y$ associé à la même valeur propre [Tenenhaus, 1998]. L'analyse inter-batterie est donc une analyse canonique où l'on remplace les métriques de Mahalanobis par les métriques identités. Cette méthode est une solution efficace au problème de multicolinéarité car ses composantes s'obtiennent sans inverser de matrice. Les solutions d'ordre suivant sont obtenues par les mêmes décompositions spectrales des matrices $(1/N)X'YY'X$ et $(1/N)Y'XX'Y$, en ajoutant les contraintes que les axes extraits sont orthogonaux aux précédents ($w^{(h)} \perp w^{(h')}$ et $v^{(h)} \perp v^{(h')}$ pour $h \neq h'$).

L'analyse de concordance proposée par Lafosse [1997] est une autre vision de l'analyse factorielle inter-batterie, où est ajoutée la notion de concordance et de discordance dans l'explication d'un tableau Y par un tableau X. L'orthogonalité des axes $(w^{(1)}, \ldots, w^{(h)})$ autorise, dimension par dimension, la décomposition de l'inertie du tableau X en parts concordantes, discordantes et de bruit, vis à vis du tableau Y [Lafosse et Hanafi, 1997] ; l'orthogonalité des axes $(v^{(1)}, \ldots, v^{(h)})$ permet d'opérer de même avec Y. La somme des contributions des concordances et discordances des

variables du tableau X avec celles du tableau Y est l'expression de leur dépendance linéaire. Le bruit est la part de Y, non explicable linéairement par les variables X.

Régression *PLS*

La régression *Partial Least Squares*, ou *PLS*, est basée sur une modification, portant sur les déflations, de l'analyse inter-batterie. Les solutions d'ordre un sont donc équivalentes. Cette méthode est initialement présentée sous forme d'un algorithme itératif, l'algorithme *NIPALS* [Wold, 1966]. L'équivalence avec la maximisation d'un critère de covariance est proposée par Höskuldsson [1988] et reprise plus en détail dans le cadre de l'algorithme *SIMPLS*, par De Jong [1993]. Les résultats issus de *NIPALS* et *SIMPLS* sont équivalents dans le cas d'une seule variable y et très proches dans le cas de plusieurs [De Jong, 1993]. La régression *PLS* est donc aussi basée sur la maximisation du critère (3.12). La solution est donnée par $w^{(1)}$, vecteur propre de la matrice $(1/N)X'YY'X$ associé à la plus grande valeur propre $\lambda^{(1)}$, et $v^{(1)}$ vecteur propre de la matrice $(1/N)Y'XX'Y$ associé à la même valeur propre.

Les composantes d'ordre suivant $(t^{(2)},\dots,t^{(h)})$ sont obtenues par la procédure de déflation de la matrice X sur les composantes $(t^{(1)},t^{(2)},\dots)$ obtenues lors des étapes précédentes, décrite dans le paragraphe 3.1.2, page 51. Les concepteurs de la méthode *PLS* ont également préconisé d'effectuer une déflation du tableau Y. Il est clair que cette dernière opération n'est pas nécessaire et ne change en rien les résultats du fait de l'orthogonalité des composantes t [Höskuldsson, 1988; Dayal et MacGregor, 1997]. La démarche utilisée permet d'élaborer un modèle de prédiction (3.13), obtenu par la régression des variables du tableau Y sur les composantes $(t^{(1)},\dots,t^{(h)})$, combinaisons linéaires des variables du tableau X.

$$\begin{aligned} Y &= t^{(1)}c^{(1)'} + \cdots + t^{(h)}c^{(h)'} + Y^{(h)} \\ t^{(1)} &= Xw^{*(1)},\ \dots,\ t^{(h)} = Xw^{*(h)} \\ \text{D'où}\quad Y &= X(w^{*(1)}c^{(1)'} + \cdots + w^{*(h)}c^{(h)'}) + Y^{(h)} \end{aligned} \tag{3.13}$$

3.2 Vision synthétique des méthodes liant X et Y

3.2.1 Uniformité des critères associés à différentes contraintes

Nous pouvons citer quelques auteurs faisant le lien entre les principales méthodes liant un tableau X à un tableau Y [Van de Geer, 1984; Chessel et Mercier, 1993; Burnham *et al.*, 1996; Tenenhaus, 1999; Rosipal et Krämer, 2006]. Van de Geer [1984] relie l'analyse canonique, l'*ACPVI* et l'analyse inter-batterie grâce à l'utilisation ou non des projecteurs des matrices X ou Y. L'utilisation de la projection de X sur l'espace des Y, ainsi que de la projection de Y sur l'espace des X mène à l'analyse canonique. L'utilisation de la projection de Y sur l'espace des X conduit à l'*ACPVI*. Le fait d'utiliser directement les matrices X et Y donne la solution de la régression *PLS*. Chessel et Mercier [1993] proposent, à travers l'analyse de co-inertie, une généralisation de ces méthodes, en se basant sur différents schémas de dualité [Cazes, 1970]. Burnham *et al.* [1996] relient eux aussi l'analyse canonique, l'*ACPVI*

(ainsi que la méthode *reduced rank regression* associée) et la régression *PLS* dans la maximisation du critère (3.14).

$$cov[(I - P^{(h-1)})Xw^{(h)}, Yv^{(h)}] \quad \text{avec} \quad P^{(h-1)} = \sum_{j=1}^{h-1} t^{(j)} t^{(j)'} / t^{(j)'} t^{(j)} \tag{3.14}$$
$$w^{'(h)} M_1 w^{(h)} = 1, \quad v^{'(h)} M_2 v^{(h)} = 1, \quad w^{(h')} M_3 w^{(h)} = 0 \quad \text{pour} \quad h' < h$$

Les contraintes utilisées varient selon les méthodes : $M_1 = M_3 = (1/N)X'X$ et $M_2 = (1/N)Y'Y$ pour l'analyse canonique, $M_1 = M_3 = (1/N)X'X$ et $M_2 = I$ pour l'*ACPVI*, $M_1 = M_2 = I$ et $M_3 = (1/N)X'X$ pour la régression *PLS*.

Tenenhaus [1999] montre que l'analyse canonique, l'analyse inter-batterie ainsi que l'*ACPVI* sont des cas particuliers de l'approche *PLS* proposée par Wold [1982], selon le mode et le schéma choisis respectivement pour déterminer le modèle de mesure et de structure. Rosipal et Krämer [2006] uniformisent le lien entre l'analyse canonique et la régression *PLS* au travers de la maximisation du critère (3.15).

$$\frac{cov^2(t,u)}{[(1-\gamma_1)var(t) + \gamma_1 I][(1-\gamma_2)var(u) + \gamma_2 I]} \quad \text{avec} \tag{3.15}$$
$$t = Xw, \quad u = Yv, \quad \|v\| = \|w\| = 1$$

A travers ce critère, les auteurs introduisent la régression *PLS* orthonormalisée, pour le cas où ($\gamma_1 = 1$) et ($\gamma_2 = 0$). Une synthèse des principales méthodes reliant deux tableaux X et Y, basée sur les critères à maximiser, ainsi que les contraintes de normes et de déflations, est donnée dans le tableau 3.1.

La majorité des méthodes liant les tableaux X et Y est basée sur la maximisation du critère $cov^2(t,u)$, auxquelles sont associées différentes contraintes de normes ou de déflations [Burnham *et al.*, 1996]. Le choix de contraintes de normes sur les composantes (cas de l'analyse canonique) oriente totalement le traitement statistique vers le lien entre X et Y, associé aux limites classiques de l'instabilité en cas de multicolinéarité des variables X et Y. Le choix de contraintes de normes sur les axes plutôt que sur les composantes (cas de la régression *PLS*) permet d'explorer des composantes liant les deux tableaux, tout en étant explicatives de ceux-ci [Burnham *et al.*, 1996]. Cette stabilité de la régression *PLS* vis à vis de la multicolinéarité des variables X et Y explique son utilisation fréquente par les praticiens de l'analyse de données. Nous retiendrons aussi l'approche de l'*ACPVI* qui oriente la description d'un tableau X vers l'explication d'un tableau Y. En posant une contrainte de norme sur la composante t plutôt que sur l'axe w, l'*ACPVI* est orientée vers l'explication des variables Y, avec l'inconvénient de sa sensibilité à la quasi-colinéarité des variables X. L'*ACPVI* optimise l'expression de l'inertie de Y expliquée par les composantes t [Burnham *et al.*, 1996]. Le choix d'avoir des axes orthogonaux (cas de l'analyse inter-batterie uniquement) oriente la méthode vers une analyse descriptive, celui d'avoir des composantes orthogonales vise à une méthode plus explicative.

3.2.2 Dimension optimale du modèle de régression

La problématique [4] (paragraphe 2.3 page 41) du traitement statistique des données d'épidémiologie animale est d'expliquer les variables Y à partir de variables

Méthode	**Critère à max.**	**Contrainte**	**Déflation**	w **vecteur propre de**
ACP / *PCR*	$\sum_p cov^2(x_p, t)$	$\|\|w\|\| = 1$	Pas de déflation ou déflation de X sur t $w\perp$ et $t\perp$	$X'X$
LRR modifiée	$\sum_q cov^2(t, y_q)$ $+\sum_p cov^2(t, x_p)$	$\|\|w\|\| = 1$	Déflation de X sur t $t\perp$	$X'YY'X + X'XX'X$
An. canonique	$cov^2(t, u)$	$\|\|t\|\| = \|\|u\|\| = 1$	Pas de déflation $t\perp$ et $u\perp$	$(X'X)^{-1}X'Y(Y'Y)^{-1}Y'X$
An. inter-batterie An. de concordance	$cov(t, u)$	$\|\|w\|\| = \|\|v\|\| = 1$	Pas de déflation $w\perp$ et $v\perp$	$X'YY'X$
Régression *PLS*	$\sum_q cov^2(t, y_q)$ ou $cov^2(t, u)$	$\|\|w\|\| = 1$ $\|\|w\|\| = \|\|v\|\| = 1$	Déflation de X et Y sur t ou déflation de X sur t $t\perp$ et $w\perp$	$X'YY'X$
ACPVI	$\sum_q cov^2(t, y_q)$ ou $cov^2(t, u)$	$\|\|t\|\| = 1$ $\|\|t\|\| = \|\|v\|\| = 1$	Pas de déflation ou déflation de X sur t $t\perp$	$(X'X)^{-1}(X'YY'X)$
PLS orthonormalisée	$cov^2(t, u)$	$\|\|w\|\| = \|\|u\|\| = 1$	Déflation de X sur t $t\perp$	$X'Y(Y'Y)^{-1}Y'X$

TAB. 3.1 – Méthodes permettant de décrire le lien entre deux tableaux X et Y.

X nombreuses et multicorrélées. Pour répondre à cette problématique, des régressions orthogonalisées sont utilisées. Quand les variables X sont quasi-colinéaires, les dernières composantes risquent d'affecter la qualité prédictive du modèle [Tomassone *et al.*, 1983; Lebart *et al.*, 2000; Barker et Brown, 2001]. Lorsque toutes les composantes ne sont pas sélectionnées pour expliquer Y, ce qui est toujours le cas en pratique, il est nécessaire de procéder par validation croisée pour sélectionner les composantes du modèle liant X à Y [Kissita, 2003, p. 28]. La validation croisée est une méthode de validation interne de modèle, initialement proposée par Stone [1974] pour l'*ACP*. Cette procédure, qui consiste à diviser le jeu de données en deux sous-échantillons, le jeu de données de calibration, permettant de déterminer les coefficients de régression et l'erreur de calibration ($RMSE_C$), et le jeu de données de validation permettant de déterminer l'erreur de validation ($RMSE_V$), est répétée m fois [Saporta, 2006, p. 426]. Cette démarche est illustrée par la figure 3.2.

L'erreur $RMSE_C$ est équivalente à l'indice $\sqrt{RSS/N}$ (=*Residual Sum of Squares*) et l'erreur $RMSE_V$ à l'indice $\sqrt{PRESS/N}$ (=*PRediction Error Sum of Squares*), calculés pour l'ensemble des variables du tableau Y [Tenenhaus, 1998, p. 83, 138]. Ces deux erreurs sont calculées par la même formule (3.16), la seule différence résidant dans le calcul de $\hat{Y}_q^{(h)}$, basé sur les individus de l'échantillon de calibration pour l'erreur $RMSE_C$, et sur les individus de l'échantillon de validation pour l'erreur $RMSE_V$.

$$RMSE^{(h)} = \|Y - \hat{Y}^{(h)}\| / \sqrt{Q} \tag{3.16}$$

Ces deux erreurs sont représentées comme des fonctions du nombre h de compo-

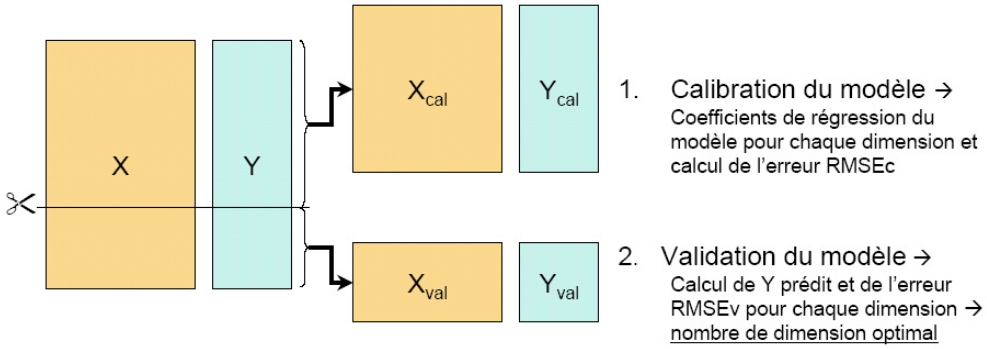


FIG. 3.2 – Validation croisée basée sur l'utilisation de deux sous-échantillons : calibration et validation.

santes $(t^{(1)},\ldots,t^{(h)})$ introduites dans le modèle de prédiction. Il est possible de définir si une composante $t^{(h)}$ améliore de façon significative la prédiction des variables Y grâce à l'indice $Q^{2(h)}$, donné pour chaque dimension du modèle $h = (1,\ldots,H)$. La composante $t^{(h)}$ a un apport significatif lorsque l'indice $Q^{2(h)} \geq 0.0975$ [Tenenhaus, 1998]. Pour le cas où les variables Y sont centrées et réduites, le calcul pour $h = 1$ de $RMSE_C^{(0)} = \sum_{q=1}^{Q}\sum_{n=1}^{N} y_{qn}^2 = \sqrt{(N-1)/N}$ [Tenenhaus, 1998, p. 83].

$$Q^{2(h)} = 1 - \frac{RMSE_V^{(h)2}}{RMSE_C^{(h-1)2}} \tag{3.17}$$

Il est de plus possible de mesurer si l'apport des h premières composantes dans l'explication des variables Y est significatif grâce à l'indice $Q_{cum}^{2(h)}$. Les composantes $(t^{(1)},\ldots,t^{(h)})$ ont un apport significatif si $Q_{cum}^{2(h)} \geq 0.5$ [Tenenhaus, 1998]. C'est cet indice qui est utilisé par la suite pour déterminer la dimension optimale h du modèle.

$$Q_{cum}^{2(h)} = 1 - \prod_{h=1}^{H} \frac{RMSE_V^{(h)2}}{RMSE_C^{(h-1)2}} \tag{3.18}$$

Chapitre 4

Continuum de méthodes permettant de décrire et relier deux tableaux

4.1 Un continuum pour cadre général aux méthodes liant deux tableaux

4.1.1 Proposition d'un continuum général

Dans le but d'explorer les liens entre les méthodes développées dans le chapitre 3, et surtout pour mieux comprendre les propriétés de ces méthodes, nous proposons de regrouper l'ensemble de ces méthodes sous un même critère à maximiser. En se référant au tableau 3.1 page 56, nous constatons que les méthodes sont basées sur des critères à maximiser proches. Un critère général permet d'unifier le critère à maximiser des trois méthodes : *ACP*, *latent root regression* modifiée et régression *PLS*. En effet, les critères associés à ces méthodes diffèrent mais les contraintes de normes et de déflations sont les mêmes. Il est donc possible de les résumer au travers de la maximisation du critère (4.1).

$$\alpha cov^2(u_\alpha^{(1)}, t_\alpha^{(1)}) + (1-\alpha)\sum_p cov^2(x_p, t_\alpha^{(1)}) \tag{4.1}$$

$$\text{avec} \quad t_\alpha^{(1)} = Xw_\alpha^{(1)}, \quad u_\alpha^{(1)} = Yv_\alpha^{(1)} \quad \|w_\alpha^{(1)}\| = \|v_\alpha^{(1)}\| = 1 \quad \text{et} \quad 0 \leq \alpha \leq 1$$

La solution est donnée par $w_\alpha^{(1)}$ vecteur propre de la matrice $(1/N^2)[\alpha(X'YY'X) + (1-\alpha)X'XX'X]$ associé à la plus grande valeur propre $\lambda_\alpha^{(1)}$. Les solutions d'ordre suivant sont obtenues par déflation de la matrice X sur les composantes t_α précédemment obtenues. On retrouve l'*ACP* pour ($\alpha = 0$), la *LRR* modifiée pour ($\alpha = 1/2$) et la régression *PLS* pour ($\alpha = 1$). De ce point de vue, la *latent root regression* modifiée apparaît comme une méthode intermédiaire entre la *PCR*, associée à l'*ACP*, et la régression *PLS*.

Les autres méthodes, analyse canonique, régression *PLS*, *ACPVI* et régression *PLS* orthogonalisée, sont basées sur la maximisation du même critère. A l'excep-

tion de l'analyse canonique, les déflations associées à ces méthodes sont équivalentes, car les solutions d'ordre supérieur à un peuvent être déterminées par la déflation de la matrice X sur les composantes t précédemment obtenues. Dans le cas de l'analyse canonique, les déflations associées aux matrices X et Y sont réalisées respectivement sur les composantes t et u. En se référant au tableau 3.1 page 56, nous constatons que seules les contraintes de normes diffèrent, ce qui a une influence sur les matrices à décomposer pour trouver les solutions d'ordre un. L'analyse canonique et la régression *PLS* sont respectivement basées sur la décomposition spectrale des matrices $[(1/N^2)(X'X)^{-1}X'Y(Y'Y)^{-1}Y'X]$ et $[(1/N^2)X'YY'X]$. De ce point de vue, il apparaît que la régression *PLS* correspond à la contraction des matrices $(X'X)^{-1}$ et $(Y'Y)^{-1}$ vers les matrices identités. Il est possible de contracter graduellement ces matrices en se basant sur la décomposition spectrale de $(1/N^2)\left[(1-\gamma_1)(X'X)+\gamma_1 I_p]^{-1}X'Y[(1-\gamma_2)(Y'Y)+\gamma_2 I_q]^{-1}Y'X\right]$ avec γ_1 et γ_2 des réels pouvant varier de 0 à 1. Cette décomposition est équivalente à la maximisation du critère (4.2).

$$\begin{aligned} &cov^2(t^{(1)},u^{(1)}) \quad \text{avec} \quad t^{(1)}=Xw^{(1)}, \quad u^{(1)}=Yv^{(1)} \\ &\gamma_1\|w^{(1)}\|^2+(1-\gamma_1)\|t^{(1)}\|^2=1 \quad \text{et} \quad \gamma_2\|v^{(1)}\|^2+(1-\gamma_2)\|u^{(1)}\|^2=1 \end{aligned} \tag{4.2}$$

Pour démontrer ce résultat, nous posons $a^{(1)}=[\gamma_1 I+(1-\gamma_1)X'X]^{1/2}w^{(1)}$ et $b^{(1)}=[\gamma_2 I+(1-\gamma_2)Y'Y]^{1/2}v^{(1)}$. Nous vérifions aisément que $\|a^{(1)}\|=\|b^{(1)}\|=1$. Ceci permet d'écrire le critère (4.2) sous la forme :

$$\begin{aligned} cov^2(t^{(1)},u^{(1)}) &= (w^{(1)'}X'Yv^{(1)})^2 \\ &= (a^{(1)'}[\gamma_1 I+(1-\gamma_1)X'X]^{-1/2}X'Y[\gamma_2 I+(1-\gamma_2)Y'Y]^{-1/2}b^{(1)})^2 \end{aligned}$$

De là, nous pouvons déduire que $a^{(1)}$ est vecteur propre de $[\gamma_1 I+(1-\gamma_1)X'X]^{-1/2}X'Y[\gamma_2 I+(1-\gamma_2)Y'Y]^{-1}Y'X[\gamma_1 I+(1-\gamma_1)X'X]^{-1/2}$, ce qui permet de conclure que $w^{(1)}$ est le vecteur propre de la matrice $[\gamma_1 I+(1-\gamma_1)X'X]^{-1}X'Y[\gamma_2 I+(1-\gamma_2)Y'Y]^{-1}Y'X$ associé à la plus grande valeur propre. De la même façon, $v^{(1)}$ est vecteur propre de la matrice $[\gamma_2 I+(1-\gamma_2)Y'Y]^{-1}Y'X[\gamma_1 I+(1-\gamma_1)X'X]^{-1}X'Y$ associé à la même valeur propre.

Les solutions d'ordre suivant sont obtenues par déflation de la matrice X sur les composantes t précédemment obtenues. Nous retrouvons l'analyse canonique pour ($\gamma_1=\gamma_2=0$), la régression *PLS* pour ($\gamma_1=\gamma_2=1$) et les solutions de l'*ACPVI* pour ($\gamma_1=0$ et $\gamma_2=1$) et de la *PLS* orthonormalisée pour ($\gamma_1=1$ et $\gamma_2=0$).

En utilisant les critères (4.1) et (4.2), il est possible de résumer l'ensemble de ces méthodes à travers la maximisation du critère (4.3) :

$$\begin{aligned} &\alpha cov^2(u^{(1)},t^{(1)})+(1-\alpha)\sum_p cov^2(x_p,t^{(1)}) \\ &\text{avec} \quad t^{(1)}=Xw^{(1)}, \quad u^{(1)}=Yv^{(1)} \\ &\gamma_1\|w^{(1)}\|^2+(1-\gamma_1)\|t^{(1)}\|^2=1, \quad \gamma_2\|v^{(1)}\|^2+(1-\gamma_2)\|u^{(1)}\|^2=1 \\ &\text{et} \quad 0\le\gamma_1\le 1, \quad 0\le\gamma_2\le 1, \quad 0\le\alpha\le 1 \end{aligned} \tag{4.3}$$

Les solutions optimales sont données par une décomposition spectrale. $w^{(1)}$ est le vecteur propre de $(1/N^2)\left[(1-\gamma_1)X'X+\gamma_1 I_p\right]^{-1}\left[\alpha X'Y[(1-\gamma_2)Y'Y+\gamma_2 I_q]^{-1}Y'X+(1-\alpha)X'XX'X\right]$

associé à la plus grande valeur propre. Les cas particuliers relatifs aux valeurs prises par les trois paramètres α, γ_1 et γ_2 sont résumés par la figure 4.1. Les méthodes classiques sont trouvées pour différentes valeurs des trois paramètres. Quand le paramètre ($\alpha = 0$), quelques soient les valeurs des paramètres γ_1 et γ_2, nous retrouvons l'*ACP* et la *PCR* associée. Quand ($\alpha = 1/2$), la version modifiée de la *LRR* est donnée pour ($\gamma_1 = \gamma_2 = 1$). Pour ($\alpha = 1$), nous retrouvons la régression *PLS* ($\gamma_1 = \gamma_2 = 1$), l'analyse canonique ($\gamma_1 = \gamma_2 = 0$), l'*ACPVI* pour ($\gamma_1 = 0$ et $\gamma_2 = 1$) et la régression *PLS* orthonormalisée pour ($\gamma_1 = 1$ et $\gamma_2 = 0$). L'intérêt d'utiliser ces trois paramètres est de pouvoir parcourir tout l'espace du cube, sans se limiter aux méthodes connues. Nous comprenons par exemple que la solution d'ordre un de la régression *PLS* est un cas particulier intermédiaire entre la solution de l'*ACP* et celle de l'*ACPVI*, mais qu'une infinité d'autres cas intermédiaires existe [Lorber *et al.*, 1987].

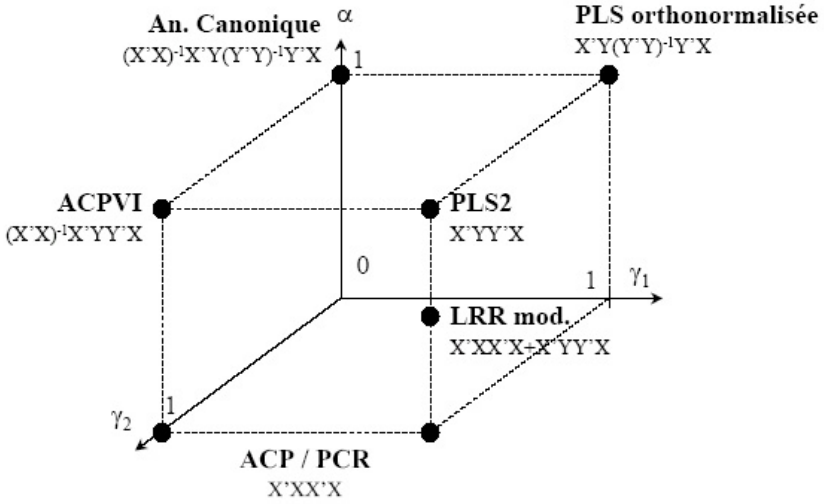


FIG. 4.1 – Illustration des cas particuliers du continuum généralisant les principales méthodes liant un tableau X à un tableau Y.

4.1.2 Interprétation des paramètres du continuum

Les variables Y n'interviennent pas dans la construction des composantes d'*ACP*. Par contre, elles interviennent au même titre que les variables X dans la construction des composantes d'ordre un de la régression *PLS* ou de l'analyse canonique. Le paramètre α mesure donc l'influence des variables Y dans la construction des composantes t. Quand α varie entre 0 à 1, les composantes t, contraintes d'être dans l'espace des variables X, s'orientent vers les variables de Y. Si l'objectif est d'expliquer X, le choix ($\alpha = 0$) est préféré, si au contraire l'objectif est de lier les tableaux X et Y, le choix de ($\alpha = 1$) est conseillé.

Lorsque toutes les variables X sont non corrélées et standardisées, nous avons $(X'X) = I_P$. Dans ce cas, les composantes issues de la régression *PLS* et de l'*ACPVI* sont identiques. Plus les variables X présentent de quasi-colinéarités, plus la matrice $(X'X)$ s'éloigne de la matrice identité et plus les solutions des deux méthodes

s'éloignent. Le paramètre γ_1, qui contrôle la contraction de la matrice $(X'X)$ vers la matrice identité, peut être vu comme un paramètre permettant de prendre en compte la multicolinéarité des variables X. Le paramètre γ_2 a la même interprétation pour les variables Y. Ceci présente souvent moins d'intérêt en pratique. En épidémiologie animale, les variables Y sont toujours moins nombreuses et moins corrélées que les variables X (paragraphe 1.3.2 page 26).

4.1.3 Comparaison à d'autres continuums

Régression *ridge*

L'instabilité des coefficients de régression en cas de multicolinéarité des variables X se traduit par une augmentation de la norme du vecteur des coefficients de régression [Tomassone *et al.*, 1983]. La régression *ridge* proposée par Hoerl et Kennard [1970] régularise la régression en imposant une norme maximum au vecteur des coefficients β, ce qui se traduit par $\beta_k = [(X'X + kI)^{-1}X'y]$ avec $k \geq 0$ [Cazes, 1975]. Pour $k = 0$, les résultats de la régression linéaire multiple sont retrouvés. De Jong et Farebrother [1994] montrent que quand $k \to \pm\infty$, les résultats sont ceux de la solution d'ordre un de la régression *PLS*. L'introduction du paramètre k stabilise l'inversion de la matrice $(X'X)$; les effets dus à la quasi-colinéarité des variables X sont progressivement éliminés, mais le biais de la régression augmente [Palm et Iemma, 1995]. Frank et Friedman [1993] montrent que la régression *ridge* est basée sur la maximisation du critère (4.4).

$$\frac{cov^2(y, t_k)}{var(t_k) + k} \quad \text{avec} \quad t_k = Xw_k, \quad \|w_k\| = 1, \quad k > 0 \tag{4.4}$$

L'extension multivariée de la régression *ridge* [Brown et Zidek, 1980] est une approche complexe basée sur les produits tensoriels.

Principal covariate regression

La méthode *principal covariate regression* (*PCovR*) est proposée par De Jong et Kiers [1992]. Elle est basée sur une maximisation pondérée de l'inertie des tableaux X et Y expliquée par des composantes t_α, orthogonales mutuellement et situées dans l'espace des variables de X. De manière plus précise, le critère (4.5) à maximiser est :

$$\alpha \sum_q cov^2(y_q, t_\alpha) + (1-\alpha) \sum_p cov^2(x_p, t_\alpha) \quad \text{avec} \quad t_\alpha = Xw_\alpha, \quad \|t_\alpha\| = 1, \quad 0 \leq \alpha \leq 1 \tag{4.5}$$

De Jong et Kiers [1992] montrent que les axes w_α sont les vecteurs propres de la matrice $(1/N^2)(X'X)^{-1}X'[\alpha YY' + (1-\alpha)XX']X$. Les cas particuliers de cette méthode sont la *PCR* pour $(\alpha = 0)$ et l'*ACPVI* pour $(\alpha = 1)$. D'après De Jong et Kiers [1992], la solution intermédiaire $(\alpha = 1/2)$ peut être comparée à la régression *PLS*. Vigneau *et al.* [2002] proposent une autre vision de la méthode *principal covariate regression*. Elle est basée sur l'*ACP* de la matrice $A = [\sqrt{\alpha}Y | \sqrt{(1-\alpha)}X]$ sous contrainte que les composantes t_α de cette *ACP* soient situées dans l'espace des variables de X et de norme unité. Les auteurs montrent que la solution est donnée par $a_\alpha = (X'X)^{1/2}w_\alpha$,

vecteur propre normé de la matrice $(1/N^2)(X'X)^{-1/2}X'AA'X(X'X)^{-1/2}$ associée à la plus grande valeur propre. En réalité, cette méthode est équivalente à la méthode *principal covariate regression*. En effet :

$$
\begin{aligned}
&(1/N^2)(X'X)^{-1/2}X'AA'X(X'X)^{-1/2}a_\alpha = \lambda_\alpha a_\alpha \\
\Leftrightarrow\ &(1/N^2)(X'X)^{-1}X'AA'X(X'X)^{-1/2}a_\alpha = \lambda(X'X)^{-1/2}a_\alpha \\
\Leftrightarrow\ &(1/N^2)(X'X)^{-1}X'AA'Xw_\alpha = \lambda w_\alpha \\
\Leftrightarrow\ &(1/N^2)(X'X)^{-1}X'[\alpha YY' + (1-\alpha)XX']w_\alpha = \lambda w_\alpha
\end{aligned}
$$

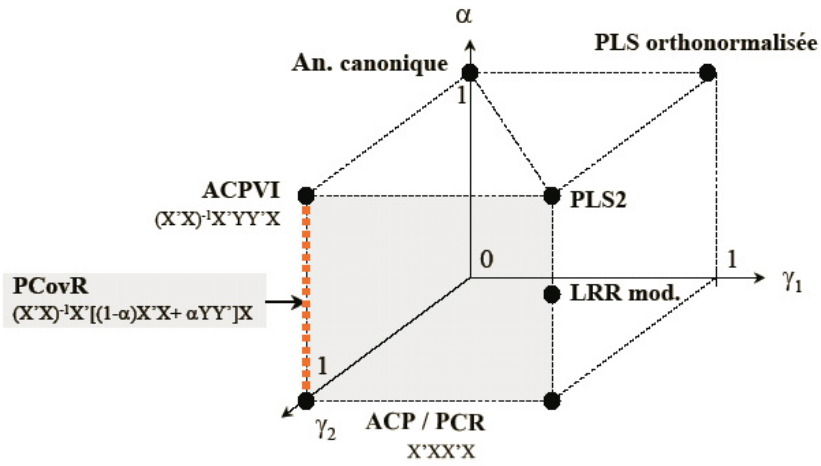


FIG. 4.2 – Illustration du domaine exploré par la méthode *principal covariate regression*.

Continuum power PLS

La méthode *continuum power PLS* proposée par De Jong *et al.* [2001] est une amélioration algorithmique de la méthode *continuum power regression* [Wise et Ricker, 1993], issue des travaux de Lorber *et al.* [1987]. L'idée de Lorber *et al.* [1987] est de substituer la matrice X par une "puissance" de X, à travers sa décomposition spectrale $X^{(\gamma)} = U\Lambda^{\gamma/2}V'$. Wise et Ricker [1993] appliquent ensuite la régression *PLS* classique à la matrice $X^{(\gamma)}$. En faisant varier le paramètre γ appliqué à la matrice des valeurs singulières, il est possible de diminuer ($\gamma < 1$) ou d'augmenter ($\gamma > 1$) le degré de multicolinéarité des variables X. Le cas particulier ($\gamma = 0$) donne la solution de la *reduced rank regression*, le cas ($\gamma = 1$) celle de la régression *PLS* et le cas ($\gamma \to +\infty$) mène à la *PCR*.

Joint continuum regression

La méthode *continuum regression* proposée par Stone et Brooks [1990] pour le cas univarié, et étendue au cas multivarié par Brooks et Stone [1994], est une méthode populaire. Dans le cas d'une seule variable y à expliquer, la méthode est basée sur la maximisation du critère $cov^2(y, t_\gamma)var(t_\gamma)^{\gamma-1}$ avec $t_\gamma = Xw_\gamma$, $\|w_\gamma\| = 1$ et $\gamma \geq 0$. Les

composantes t_γ d'ordre suivant sont obtenues par déflation des variables sur les composantes t_γ précédemment obtenues. Les cas particuliers de cette méthode sont la régression linéaire multiple ($\gamma = 0$), la régression *PLS*1 ($\gamma = 1$) et la *PCR* ($\gamma \to \infty$). Sundberg [1993] montre que lorsque γ varie entre 0 à 1, l'axe w s'oriente vers les directions associées aux plus grandes valeurs propres de la matrice $(X'X)$. Dans le cas multivarié où $Y = [y_1, \ldots, y_Q]$, le critère de la méthode *joint continuum regression* est basé sur la maximisation du critère (4.6).

$$\sum_q cov^2(y_q, t_\gamma) var(t_\gamma)^{\gamma-1} \quad \text{avec} \quad t_\gamma = Xw_\gamma, \quad \|w_\gamma\| = 1, \quad \gamma \geq 0 \tag{4.6}$$

La résolution de cette maximisation admet trois cas particuliers : la méthode *reduced rank regression* issue de l'*ACPVI* des tableaux X et Y ($\gamma = 0$) [Burnham *et al.*, 1999], la régression *PLS* ($\gamma = 1$) et l'*ACP* suivie de la régression sur composantes principales ($\gamma \to +\infty$).

Les deux continuums, *continuum power PLS* et *joint continuum regression*, explorent des espaces compris entre les mêmes cas particuliers (*ACPVI*, *PLS* et *PCR*), mais sans emprunter nécessairement les mêmes chemins. Une illustration possible est proposée par la figure 4.3.

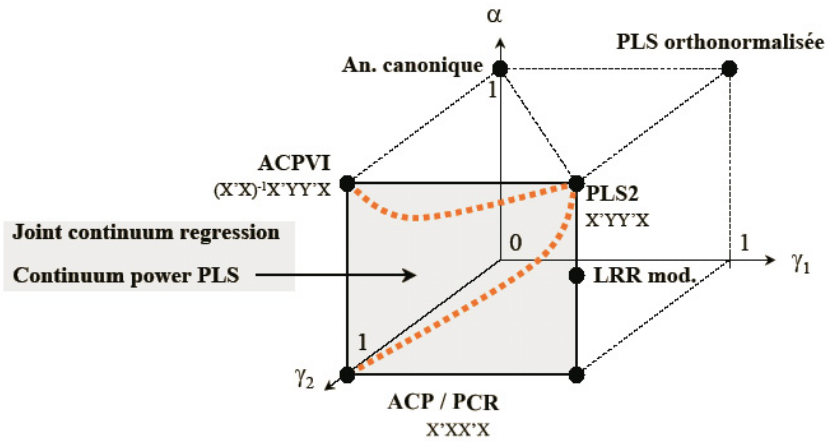


FIG. 4.3 – Illustration des domaines (possibles) explorés par les méthodes *continuum power PLS* et *joint continuum regression*.

Analyse canonique *ridge*

Vinod [1976] introduit le concept d'analyse canonique *ridge* pour stabiliser les coefficients issus de l'analyse canonique en cas de multicolinéarité des variables X et Y. Cette méthode est une extension directe de la régression *ridge* pour le cas de plusieurs variables à expliquer [De Bie *et al.*, 2005]. Rosipal et Krämer [2006] introduisent l'analyse canonique *ridge* en considérant la maximisation du critère

(4.7).

$$\frac{cov^2(t,u)}{[(1-\gamma_1)var(t)+\gamma_1][(1-\gamma_2)var(u)+\gamma_2]} \quad \text{avec} \quad t=Xw, \quad u=Yv, \quad \|w\|=\|v\|=1 \quad \text{et} \quad 0\leq\gamma_1\leq 1, \quad 0\leq\gamma_2\leq 1 \tag{4.7}$$

La solution est donnée par la décomposition spectrale de la matrice $(1/N^2)[(1-\gamma_1)X'X+\gamma_1 I_P]^{-1}X'Y[(1-\gamma_2)Y'Y+\gamma_2 I_Q]^{-1}Y'X$. D'après les résultats précédents (paragraphe 4.1.1 page 59), il est possible d'en déduire que l'analyse canonique *ridge* est basée sur la maximisation du critère $cov^2(t,u)$ sous les contraintes de norme $\gamma_1\|w\|^2+(1-\gamma_1)\|t\|^2=1$ et $\gamma_2\|v\|^2+(1-\gamma_2)\|u\|^2=1$ avec $0\leq\gamma_1\leq 1$ et $0\leq\gamma_2\leq 1$. Le cas particulier de l'analyse canonique est retrouvé pour ($\gamma_1=\gamma_2=0$), celui de la régression *PLS* pour ($\gamma_1=\gamma_2=1$) et de la méthode *PLS* orthonormalisée pour ($\gamma_1=1$ et $\gamma_2=0$). On peut noter de plus que le cas ($\gamma_1=0$ et $\gamma_2=1$) est celui de l'*ACPVI* (non précisé par les auteurs). La figure 4.4 illustre le domaine exploré par le double continuum de l'analyse canonique *ridge*.

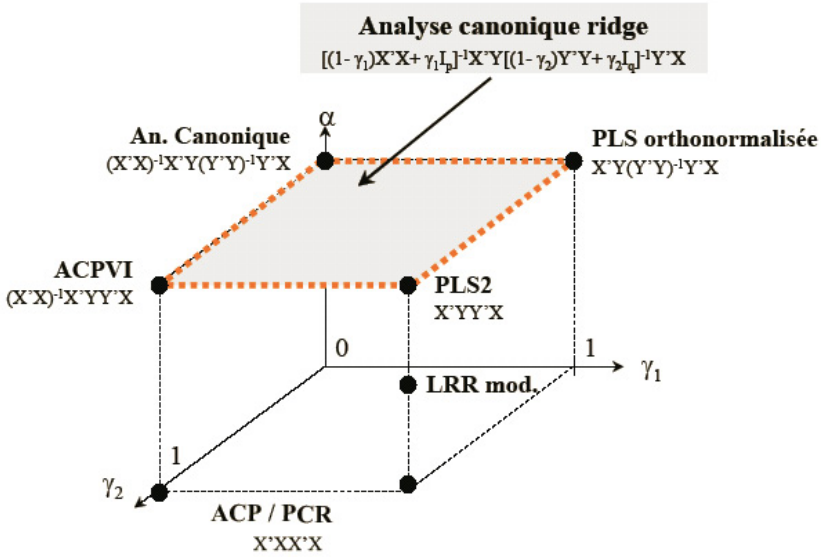


FIG. 4.4 – Illustration du domaine exploré par l'analyse canonique *ridge*.

4.1.4 Sélection des continuums à explorer dans le cadre du traitement des données d'épidémiologie animale

L'idée d'un continuum général est intéressante du point de vue théorique. Chaque paramètre a un rôle clair qui facilite son interprétation. Le continuum permet d'ajuster à la fois à l'objectif du traitement statistique (paramètre α) et de mieux apréhender le problème de la multicolinéarité des variables X et Y (paramètres γ_1 et γ_2). L'utilisation conjointe de ces trois paramètres permet d'explorer l'ensemble des méthodes comprises dans le cube présenté paragraphe 4.1 page 61, admettant des cas particuliers connus. Il faut noter de plus qu'un quatrième paramètre d'ajustement s'ajoute aux précédents : le nombre de dimension h du modèle liant X à

Y. Cependant, il semble peu raisonnable d'utiliser un continuum dont il faut déterminer quatre paramètres, à partir de tableaux de données ne contenant qu'une centaine d'individus.

Les continuums explorés doivent répondre aux problématiques relatives aux données d'épidémiologie animale exposées dans le paragraphe 2.3 page 41. La problématique [1], principale préoccupation du traitement statistique de ce type de données, oriente vers des continuums permettant la prise en compte de la quasi-colinéarité des variables X. Un continuum lié à la variation du paramètre γ_1 est donc privilégié. La prise en compte de la quasi-colinéarité des variables Y, vue au travers des variations du paramètre γ_2, est une voie non prioritaire qui ne sera pas explorée. Les objectifs de ces traitements étant à la fois basés sur l'explication des variables X et sur leurs liens avec les variables Y, un continuum basé sur la variation du paramètre α est aussi exploré. De plus, les continuums doivent être facilement étendus à la problématique multibloc [3], détaillée par la suite dans le chapitre 7 page 111.

4.2 Continuums explorés dans le cadre de deux tableaux

4.2.1 Continuum *latent root regression*

Le premier continuum exploré est lié à l'objectif du traitement statistique des données, orienté vers l'explication des variables X ou Y. Ce continuum est basé sur la variation du paramètre α. Il est choisi pour sa robustesse à la quasi-colinéarité des variables X (paramètre $\gamma_1 = 1$). L'*ACP*, la *latent root regression* modifiée et la régression *PLS* sont trois méthodes intéressantes du point de vue de la description des variables X, plus ou moins orientée vers des variables à expliquer Y. Elles sont de plus relativement robustes à la quasi-colinéarité entre les variables X. Le choix est fait de définir une famille de méthodes englobant ces trois cas particuliers. Ce continuum est basé sur la maximisation du critère (4.8).

$$(1-\alpha)\sum_p cov^2(x_p, t_\alpha^{(1)}) + \alpha \sum_q cov^2(y_q, t_\alpha^{(1)}) \tag{4.8}$$
$$\text{avec} \quad t_\alpha^{(1)} = X w_\alpha^{(1)}, \quad \|w_\alpha^{(1)}\| = 1 \quad \text{et} \quad 0 \leq \alpha \leq 1$$

La solution est donnée par $w_\alpha^{(1)}$ premier vecteur propre de la matrice $(1/N^2)[\alpha X'YY'X + (1-\alpha)X'XX'X]$ associé à la plus grande valeur propre $\lambda_\alpha^{(1)}$. La solution d'ordre suivant est obtenue par remplacement de la matrice X par son résidu de la régression sur la première composante $t_\alpha^{(1)}$ dans le critère à maximiser. La figure 4.5 illustre la famille de méthodes explorées par le continuum *LRR*. En faisant varier le paramètre α, le continuum explore les solutions comprises entre l'*ACP* de X ($\alpha = 0$), la version modifiée de la *latent root regression* ($\alpha = 1/2$) et la régression *PLS* ($\alpha = 1$).

La maximisation du critère $\sum_p cov^2(x_p, t_\alpha) + \sum_q cov^2(y_q, t_\alpha)$ revient à maximiser le critère équivalent $\sum_j cov^2(a_j, t_\alpha)$, où a_j désigne la variable générique du tableau $A = [Y|X]$, avec $t_\alpha = X w_\alpha$ et $\|w_\alpha\| = 1$. Ce critère conduit donc à une régression *PLS* du tableau concaténé $A = [Y|X]$ sur le tableau X. En intégrant la pondération dans le tableau initial $A_\alpha = [\sqrt{\alpha}Y|\sqrt{(1-\alpha)}X]$, nous retrouvons la solution du continuum *LRR*

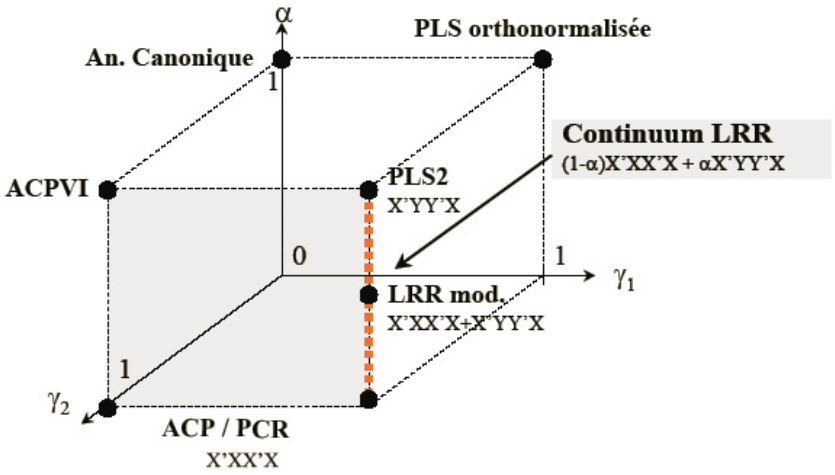


FIG. 4.5 – Illustration du domaine exploré par le continuum *LRR*.

grâce à la décomposition de la matrice $(1/N^2)X'A_\alpha A'_\alpha X = (1/N^2)(X'[\alpha YY' + (1-\alpha)XX']X) = (1/N^2)[\alpha X'YY'X + (1-\alpha)X'XX'X]$.

Propriétés du continuum *LRR*

D'après le critère (4.8), il apparaît intuitivement que lorsque le paramètre α augmente, la solution du problème est davantage orientée vers l'explication des variables Y. Nous allons étayer cette idée en démontrant que la fonction $\sum_q cov^2(y_q, t_\alpha^{(1)})$ est une fonction croissante du paramètre α. Nous considérons deux valeurs du paramètre α, α_1 et α_2, telles que $0 \leq \alpha_1 \leq \alpha_2 \leq 1$. Soit les vecteurs w_{α_1} et w_{α_2}, associés respectivement aux composantes t_{α_1} et t_{α_2}, solutions du problème d'optimisation (4.8) vérifiant :

$$(1-\alpha_1)\sum_p cov^2(x_p, t_{\alpha_2}) + \alpha_1 \sum_q cov^2(y_q, t_{\alpha_2}) \leq (1-\alpha_1)\sum_p cov^2(x_p, t_{\alpha_1}) + \alpha_1 \sum_q cov^2(y_q, t_{\alpha_1})$$
$$(1-\alpha_2)\sum_p cov^2(x_p, t_{\alpha_1}) + \alpha_2 \sum_q cov^2(y_q, t_{\alpha_1}) \leq (1-\alpha_2)\sum_p cov^2(x_p, t_{\alpha_2}) + \alpha_2 \sum_q cov^2(y_q, t_{\alpha_2})$$

Ce qui permet d'écrire :

$$\sum_p \left[cov^2(x_p, t_{\alpha_2}) - cov^2(x_p, t_{\alpha_1})\right] \leq \frac{\alpha_1}{1-\alpha_1} \sum_q \left[cov^2(y_q, t_{\alpha_1}) - cov^2(y_q, t_{\alpha_2})\right] \quad (4.9)$$
$$\sum_p \left[cov^2(x_p, t_{\alpha_1}) - cov^2(x_p, t_{\alpha_2})\right] \leq \frac{\alpha_2}{1-\alpha_2} \sum_q \left[cov^2(y_q, t_{\alpha_2}) - cov^2(y_q, t_{\alpha_1})\right] \quad (4.10)$$

Nous allons montrer que $\sum_q cov^2(y_q, t_\alpha^{(1)})$ est une fonction croissante de α. Comme la fonction $f(\alpha) = \alpha/(1-\alpha)$ est une fonction croissante de α, nous avons $\alpha_1/(1-\alpha_1) \leq \alpha_2/(1-\alpha_2)$. En faisant l'hypothèse qu'il existe α_1 et α_2 tels que $\sum_q cov^2(y_q, t_{\alpha_1}) >$

$\sum_q cov^2(y_q, t_{\alpha_2})$, nous obtenons :

$$\frac{\alpha_1}{1-\alpha_1}\sum_q \left[cov^2(y_q, t_{\alpha_1}) - cov^2(y_q, t_{\alpha_2})\right] \leq \frac{\alpha_2}{1-\alpha_2}\sum_q \left[cov^2(y_q, t_{\alpha_1}) - cov^2(y_q, t_{\alpha_2})\right] \tag{4.11}$$

En combinant les inégalités (4.9) et (4.11), nous obtenons :

$$\sum_p \left[cov^2(x_p, t_{\alpha_2}) - cov^2(x_p, t_{\alpha_1})\right] \leq \frac{\alpha_2}{1-\alpha_2}\sum_q \left[cov^2(y_q, t_{\alpha_1}) - cov^2(y_q, t_{\alpha_2})\right]$$
$$\Leftrightarrow \sum_p \left[cov^2(x_p, t_{\alpha_1}) - cov^2(x_p, t_{\alpha_2})\right] \geq \frac{\alpha_2}{1-\alpha_2}\sum_q \left[cov^2(y_q, t_{\alpha_2}) - cov^2(y_q, t_{\alpha_1})\right]$$

Cette dernière inégalité est en contradiction avec l'inégalité (4.10). Nous en déduisons que $\sum_q cov^2(y_q, t_\alpha)$ est fonction croissante de α. De manière simulaire, nous pouvons montrer que $\sum_p cov^2(x_p, t_\alpha)$ est fonction décroissante de α.

Comparaison à d'autres continuums

Il est aisé de voir que le continuum *LRR* parcourt un chemin parallèle à la méthode *principal covariate regression* (paragraphe 4.1.3 page 62). En effet, ces deux méthodes sont basées sur le même critère à maximiser à la seule différence des contraintes de norme : $\|t\| = 1$ pour *PCovR* et $\|w\| = 1$ pour le continuum *LRR*. On note que les solutions d'ordre supérieur à un de *PCovR* peuvent s'obtenir par déflation de la matrice X sur les composantes t précédemment obtenues (même démonstration que pour l'*ACPVI*, présentée paragraphe 3.1.1 page 47), ce qui uniformise le mode de calcul des deux méthodes. Cette différence de contrainte de norme fait que le continuum *LRR* est moins orienté vers l'explication des variables de Y que *PCovR*, et est moins sensible à la quasi-colinéarité des variables de X. En reprenant la vision de Vigneau *et al.* [2002], il est aussi possible d'affirmer que les solutions de la méthode *PCovR* sont obtenues par l'*ACPVI* des tableaux $[\sqrt{\alpha}Y|\sqrt{(1-\alpha)}X]$ et X, et que celles du continuum *LRR* par la régression *PLS* de $[\sqrt{\alpha}Y|\sqrt{(1-\alpha)}X]$ sur X.

4.2.2 Continuum $ACPVI - PLS$ regression

Dans le cas d'un tableau X orienté vers l'explication d'un tableau Y, le choix de la méthode diffère selon le degré de multicolinéarité des variables X. Si celles-ci sont peu corrélées entre elles, le choix d'une méthode orientée vers la prédiction des variables Y, telle que l'*ACPVI*, est recommandé. Si les variables, comme c'est souvent le cas, comportent des quasi-colinéarités, le choix se porte sur une méthode plus stable comme la régression *PLS*. Ces deux méthodes sont respectivement basées sur la décomposition spectrale des matrices $[(1/N^2)(X'X)^{-1}X'YY'X]$ et $[(1/N^2)X'YY'X]$. Il apparaît donc que la régression *PLS* est basée sur une contraction de la matrice $(X'X)^{-1}$ vers la matrice identité I_P. Le continuum détaillé ici établit un lien entre l'*ACPVI* et la régression *PLS*. Nous nous y référerons par l'appellation continuum

ACPVI – PLS. Il est basé sur la variation du paramètre γ_1. Comme nous l'avons déjà souligné, cette stratégie est basée sur la maximisation du critère (4.12).

$$\sum_q cov^2(y_q, t_{\gamma_1}^{(1)}) \quad \text{avec} \quad t_{\gamma_1}^{(1)} = X w_{\gamma_1}^{(1)} \tag{4.12}$$
$$(1-\gamma_1)\|t_{\gamma_1}^{(1)}\|^2 + \gamma_1 \|w_{\gamma_1}^{(1)}\|^2 = 1 \quad \text{et} \quad 0 \leq \gamma_1 \leq 1$$

La solution de ce problème de maximisation est donnée par $w_{\gamma_1}^{(1)}$, vecteur propre de la matrice $(1/N^2)[(1-\gamma_1)(X'X)+\gamma_1 I]^{-1}X'YY'X$ associé à la plus grande valeur propre $\lambda_{\gamma_1}^{(1)}$. Les composantes d'ordre suivant $(t_{\gamma_1}^{(1)}, \ldots, t_{\gamma_1}^{(h)})$ sont obtenues en remplaçant dans le critère (4.12), X par son résidu de la régression sur les composantes précédemment obtenues. La figure 4.6 illustre la famille de méthodes explorée par le continuum *ACPVI – PLS*. En faisant varier le paramètre γ_1, le continuum explore les solutions comprises entre celle de l'*ACPVI* ($\gamma_1 = 0$) et de la régression *PLS* ($\gamma_1 = 1$). Les résultats de l'*ACPVI* sont orientés vers l'explication des variables Y du fait de la contrainte $\|t\| = 1$. Par contre ils présentent plus de sensibilité à la quasi-colinéarité des variables X, car ils nécessitent l'inversion de la matrice $(X'X)$. Le risque est de trouver des variables latentes t de faible variance (liées au bruit) ayant un bon pouvoir explicatif des variables Y. A l'inverse, les résultats de la régression *PLS* sont moins orientés vers l'explication des variables Y car les composantes t sont construites de façon à expliquer correctement les variables X ($\|w\| = 1$ et non pas $\|t\| = 1$). Ils présentent peu de sensibilité à la multicolinéarité des variables X car la matrice $(X'X)$ n'a pas besoin d'être inversée. Cependant, cette solution plus stable risque de déterminer des composantes t de forte variance expliquant peu Y.

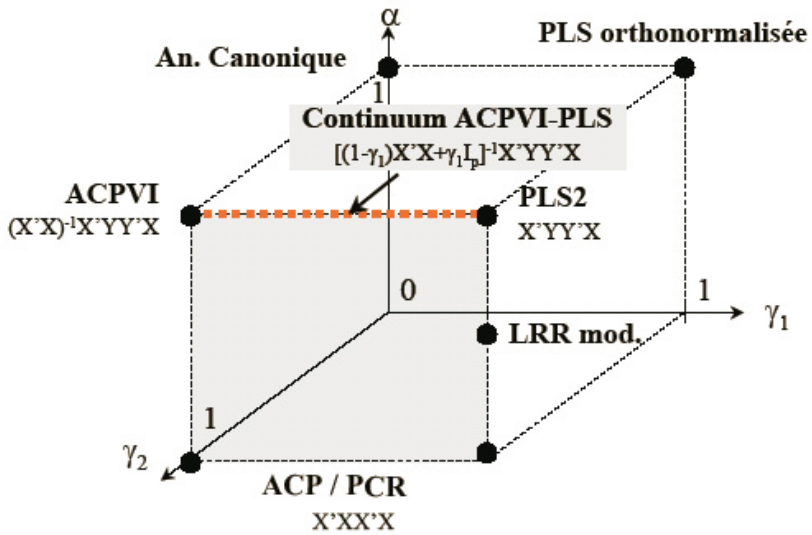


FIG. 4.6 – Illustration du domaine exploré par le continuum *ACPVI – PLS*.

L'approche continuum proposée peut être introduite à partir de la maximisation

du citère Q_{γ_1} :

$$Q_{\gamma_1} = \frac{w'_{\gamma_1} X'YY'Xw_{\gamma_1}}{w'_{\gamma_1}[(1-\gamma_1)X'X+\gamma_1 I]w_{\gamma_1}}$$

$$Q_{\gamma_1} = \frac{t'_{\gamma_1} YY't_{\gamma_1}}{(1-\gamma_1)t'_{\gamma_1}t_{\gamma_1}+\gamma_1.w'_{\gamma_1}w_{\gamma_1}}$$

Cette expression étant invariante par multiplication par un facteur d'échelle, nous imposons arbitrairement $\|w_{\gamma_1}\| = 1$, ce qui permet d'écrire le critère Q_{γ_1} sous la forme (4.13).

$$\frac{\sum_q cov^2(y_q, t_{\gamma_1})}{(1-\gamma_1)var(t_{\gamma_1})+\gamma_1} \tag{4.13}$$

Cette nouvelle expression du critère va nous permettre de démontrer des propriétés intéressantes liées à ce continuum et d'établir un lien entre ce continuum et d'autres stratégies d'analyse.

Propriétés du continuum $ACPVI-PLS$

Sensibilité à la multicolinéarité La sensibilité du modèle à la multicolinéarité des variables X peut être reflétée par l'indice de conditionnement η [Belsley *et al.*, 1980]. L'indice de conditionnement maximal est le rapport de la plus grande valeur propre de la matrice $(X'X)$, $\lambda^{(1)}$ sur la plus petite, $\lambda^{(P)}$ [Erkel-Rousse, 1995]. Une grande valeur de η alerte sur la présence de quasi-colinéarité au sein des variables X et donc sur le risque d'instabilité du modèle. D'après Erkel-Rousse [1995], les liaisons fortes entre les variables X sont associées à des valeurs du paramètre η de l'ordre de 30 (situation de risque entre 20 et 30). L'indice de conditionnement maximal de la matrice $[(1-\gamma_1)X'X+\gamma_1 I]$ est donné par :

$$\eta_{\gamma_1} = \frac{(1-\gamma_1)\lambda^{(1)}+\gamma_1}{(1-\gamma_1)\lambda^{(P)}+\gamma_1}$$

Il est aisé de montrer, en étudiant la dérivée de cette fonction par rapport à γ_1, que l'indice η_{γ_1} est une fonction décroissante du paramètre γ_1. Au sein du continuum $ACPVI-PLS$, la régression PLS correspond donc à la plus petite valeur de η_{γ_1}, l'$ACPVI$ à la plus grande.

Stabilité du modèle Comme nous l'avons déjà souligné, la stabilité du modèle obtenu par la régression des variables Y sur la première composante $t^{(1)}_{\gamma_1}$ peut être évaluée par la variance de $t^{(1)}_{\gamma_1}$. Nous allons montrer que $var(t^{(1)}_{\gamma_1})$ est une fonction croissante de γ_1, ce qui signifie que lorsque le paramètre γ_1 augmente, le continuum s'éloigne des composantes de faible variance reflétant le bruit pour investir des directions plus stables. La solution d'ordre un de l'$ACPVI$ correspond à la solution de plus faible variance de la composante $t^{(1)}_{\gamma_1}$ et la régression PLS à la solution de plus grande variance. Afin de démontrer cette propriété, considérons deux valeurs du paramètre γ_1, $\gamma_{1(1)}$ et $\gamma_{1(2)}$, telles que $0 \leq \gamma_{1(1)} \leq \gamma_{1(2)} \leq 1$. Soit les vecteurs $w_{\gamma_{1(1)}}$ et

$w_{\gamma_{1(2)}}$, solutions du problème d'optimisation (4.13), en remplaçant γ_1 respectivement par $\gamma_{1(1)}$ et $\gamma_{1(2)}$:

$$\frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(2)}})}{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(1)}} \leq \frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(1)}})}{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(1)}})+\gamma_{1(1)}} \tag{4.14}$$

$$\frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(1)}})}{(1-\gamma_{1(2)})var(t^{(1)}_{\gamma_{1(1)}})+\gamma_{1(2)}} \leq \frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(2)}})}{(1-\gamma_{1(2)})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(2)}} \tag{4.15}$$

La multiplication terme à terme des inégalités (4.14) et (4.15) donne après simplification :

$$\frac{(1-\gamma_{1(2)})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(2)}}{(1-\gamma_{1(2)})var(t^{(1)}_{\gamma_{1(1)}})+\gamma_{1(2)}} \leq \frac{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(1)}}{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(1)}})+\gamma_{1(1)}} \tag{4.16}$$

Après simplification de l'inégalité (4.16), il ressort que $var(t^{(1)}_{\gamma_{1(1)}}) \leq var(t^{(1)}_{\gamma_{1(2)}})$, ce qui permet de conclure que $var(t^{(1)}_{\gamma_1})$ est une fonction croissante de γ_1.

Qualité d'ajustement du modèle La qualité d'ajustement du modèle obtenu en régressant les variables Y sur la première composante $t^{(1)}_{\gamma_1}$ est donnée par l'inertie de Y expliquée par cette composante, $Iner\_Expl(Y, t^{(1)}_{\gamma_1}) = \sum_q cov^2(y_q, t^{(1)}_{\gamma_1})/var(t^{(1)}_{\gamma_1})$. Nous allons montrer que cette inertie expliquée est une fonction décroissante du paramètre γ_1. Il s'ensuit qu'au sein du continuum $ACPVI-PLS$, l'$ACPVI$ correspond à la solution de plus grande capacité d'ajustement du modèle et la régression PLS à la plus faible. Comme précédemment, nous considérons deux valeurs du paramètre γ_1, $\gamma_{1(1)}$ et $\gamma_{1(2)}$, telles que $0 \leq \gamma_{1(1)} \leq \gamma_{1(2)} \leq 1$. $var(t^{(1)}_{\gamma_1})$ étant une fonction croissante de γ_1, $var(t^{(1)}_{\gamma_{1(1)}}) \leq var(t^{(1)}_{\gamma_{1(2)}})$, il s'ensuit :

$$(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(1)}})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(1)}var(t^{(1)}_{\gamma_{1(1)}}) \leq (1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(1)}})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(1)}var(t^{(1)}_{\gamma_{1(2)}})$$

Ce qui est équivalent à :

$$\frac{var(t^{(1)}_{\gamma_{1(1)}})}{var(t^{(1)}_{\gamma_{1(2)}})} \leq \frac{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(1)}})+\gamma_{1(1)}}{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(1)}}$$

De l'inégalité (4.14) précédente, nous déduisons que :

$$\frac{var(t^{(1)}_{\gamma_{1(1)}})}{var(t^{(1)}_{\gamma_{1(2)}})} \leq \frac{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(1)}})+\gamma_{1(1)}}{(1-\gamma_{1(1)})var(t^{(1)}_{\gamma_{1(2)}})+\gamma_{1(1)}} \leq \frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(1)}})}{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(2)}})}$$

$$\text{Soit} \quad \frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(2)}})}{var(t^{(1)}_{\gamma_{1(2)}})} \leq \frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_{1(1)}})}{var(t^{(1)}_{\gamma_{1(1)}})}$$

Ce qui montre que l'inertie de Y expliquée par la première composante $t^{(1)}_{\gamma_1}$, $Iner\_Expl(Y, t^{(1)}_{\gamma_1}) = \sum_q cov^2(y_q, t^{(1)}_{\gamma_1})/var(t^{(1)}_{\gamma_1})$ est une fonction décroissante de γ_1.

Norme du vecteur des coefficients de régression Dans le cas où la régression *OLS* (=*Ordinary Least Square*) classique est appliquée sur des variables X quasi-colinéaires, la norme des coefficients de la régression peut être artificiellement élevée. Les méthodes de régression telles que la régression *PLS* permettent de réduire la norme des coefficients de régression et par conséquent de régulariser les coefficients de régression. Pour le cas univarié, $Y = [y]$, De Jong [1995] démontre que la norme des coefficients de la régression *PLS*1 est réduite, en comparaison à celle des coefficients de la régression *OLS*. Il est connu que la *reduced rank regression*, réalisée en régressant les variables Y sur les composantes de l'*ACPVI*, est une extension de la régression *OLS* pour le cas multivarié $Y = [y_1, \ldots, y_Q]$. Nous démontrons que dans le cas d'un modèle avec une seule composante issue du continuum $ACPVI - PLS$, la norme du vecteur de coefficients $\|\beta^{(1)}_{\gamma_1}\|$ est une fonction décroissante du paramètre γ_1. Dans le cas d'un modèle à une seule dimension, le modèle de prédiction est obtenu en régressant chaque variable Y sur $t^{(1)}_{\gamma_1}$, ce qui mène au modèle $\hat{Y}^{(1)}_{\gamma_1} = X\beta^{(1)}_{\gamma_1}$. Le carré de la norme du vecteur $\beta^{(1)}_{\gamma_1}$ est donné par :

$$\|\beta^{(1)}_{\gamma_1}\|^2 = \frac{\sum_q cov^2(y_q, t^{(1)}_{\gamma_1})}{var^2(t^{(1)}_{\gamma_1})} = \frac{Iner\_Expl(Y, t^{(1)}_{\gamma_1})}{var(t^{(1)}_{\gamma_1})}$$

Cette fonction étant le rapport d'une fonction qui décroît et d'une fonction qui croît avec le paramètre γ_1, nous en déduisons que $\|\beta^{(1)}_{\gamma_1}\|^2$ est une fonction décroissante de γ_1. En particulier, nous en déduisons que la norme du vecteur des coefficients d'ordre un de la régression *PLS* ($\gamma_1 = 1$) est plus petite que celle de l'*ACPVI* ($\gamma_1 = 0$).

Comparaison à d'autres continuums

En appliquant le continuum $ACPVI - PLS$ au cas univarié $Y = [y]$, nous recherchons le premier vecteur propre de la matrice $[(1-\gamma_1)X'X + \gamma_1 I]^{-1} X'yy'X$, ou de manière équivalente $\left[X'X + k_{\gamma_1} I\right]^{-1} X'yy'X$ avec $k_{\gamma_1} = \gamma_1/(1-\gamma_1)$. Cette dernière matrice est de rang un, et le premier vecteur propre associé à cette matrice est donné par $w^{(1)}_{\gamma_1} = \left[X'X + k_{\gamma_1} I\right]^{-1} X'y$. Le modèle obtenu en régressant la variable à expliquer y sur la première composante $t^{(1)}_{\gamma_1} = Xw^{(1)}_{\gamma_1}$ est donné par $\hat{y}^{(1)}_{\gamma_1} = aX\left[X'X + k_{\gamma_1} I\right]^{-1} X'y$, a étant un scalaire. Ceci montre que dans le cas univarié, le continuum $ACPVI - PLS$ est directement lié à la régression *ridge* [Hoerl et Kennard, 1970]. Cette stratégie d'analyse est discutée plus en détail par Qannari et Hanafi [2005]. Ces auteurs se réfèrent à cette méthode, sous le nom de *simple continuum regression*. Ils montrent sur un exemple et sur la base d'une validation croisée, qu'en plus de sa simplicité, cette méthode donne de meilleurs résultats que d'autres méthodes plus complexes comme le continuum de Stone et Brooks [1990]. Les autres avantages du continuum $ACPVI - PLS$ sont tout d'abord d'apporter des composantes supplémentaires permettant d'améliorer la qualité de prédiction de la variable y, mais aussi de pouvoir être étendue directement au cas de plusieurs variables Y à expliquer comme nous l'avons fait ci-dessus.

4.2.3 Sélections des paramètres optimaux des continuums

Quelque soit le continuum exploré, deux paramètres doivent être fixés : le paramètre du continuum, compris entre 0 et 1 (α ou γ_1 selon le continuum étudié), ainsi que le nombre h de composantes $(t^{(1)}, \ldots, t^{(h)})$ à retenir dans le modèle expliquant Y par X. La procédure de validation croisée [Stone, 1974], ainsi que les calculs associés, détaillés dans le paragraphe 3.2.2 page 55, sont utilisés à l'identique. La seule différence réside dans le fait que le schéma 3.2 est appliqué pour chaque valeur du paramètre du continuum. Pour une dimension donnée, les erreurs de validation $RMSE_V$ (équation (3.16) page 56) associées à une gamme de valeurs du paramètre du continuum, obtenues par discrétisation entre 0 et 1, sont comparées. C'est la valeur du paramètre associée à la plus petite erreur de validation $RMSE_V$ qui est choisie comme étant la valeur optimale du continuum ($\alpha_{opt}^{(h)}$ ou $\gamma_{1,opt}^{(h)}$). L'utilisation de la significativité des indices Q^2 et Q^2_{cum} (équations (3.17) et (3.18) page 57) permet d'utiliser les informations conjointes des deux erreurs, $RMSE_C$ et $RMSE_V$ pour déterminer la dimension optimale h du modèle.

Chapitre 5

Application au traitement de données d'épidémiologie animale organisées en deux tableaux

5.1 Données et problématique

Les données étudiées sont extraites de l'enquête analytique sur les facteurs de risque de la consommation d'antibiotiques en élevages de dindes, citée paragraphe 1.3.2 page 26 [Chauvin *et al.*, 2005]. L'objectif général de ce travail n'est pas d'expliquer un phénomène de santé mais de déceler les éléments, relevés en élevage, permettant de prédire le statut à risque d'un lot de dinde à l'abattoir. L'objectif du jeu de données extrait de cette enquête est de déterminer les facteurs influençant les pertes de l'éleveur, résumées par la mortalité en élevage et le taux de saisie des carcasses à l'abattoir (tableau Y contenant deux variables). Le tableau X explicatif est composé de 19 variables relatives aux caractéristiques d'élevage du lot de dindes étudié et à la performance technique et économique de ce lot. Ces variables sont décrites dans le tableau 5.1. Ces variables ayant des échelles de mesure différentes, il est nécessaire de les centrer et les réduire. Ces variables sont majoritairement quantitatives ; les six variables qualitatives, toutes codées en deux modalités, sont traitées conjointement aux autres sous forme de variables dichotomiques. Cette enquête est réalisée sur 659 lots de dindes.

Les méthodes développées pouvant être sensibles à la multicolinéarité des variables explicatives notamment, l'indice de conditionnement maximal (décrit dans le paragraphe 4.2.2 page 70) relatif à chacun des deux tableaux est calculé. L'indice de conditionnement maximal relatif au tableau X est de 638, celui relatif au tableau Y est de 1.5. On peut donc noter que le tableau des variables explicatives comporte des quasi-colinéarités majeures, car l'indice est supérieur au seuil de 30 [Erkel-Rousse, 1995].

Bloc	Variable	Description
Y	SAISIE	Pourcentage de dindes saisies à l'abattoir
	PERTE	Pourcentage de mortalité en élevage
X	FVO	Utilisation d'aliments exempts farine de viande et d'os (1=oui, 0=non)
	COUTDES	Coûts de désinfection
	MOYVET	Coût moyen des frais vétérinaires pour les trois derniers lots
	DESSER	Déserrage des animaux (1=oui, 0=non)
	DVID	Durée du vide sanitaire avant l'entrée des poussins (en *jour*)
	NBDEP	Nombre de départs à l'abattoir pour le lot étudié
	ESP	Lot précédent de la même espèce (1=oui, 0=non)
	PBSAN	Problème sanitaire sérieux durant la période d'élevage (1=oui, 0=non)
	DESINF	Type de main d'oeuvre pour la désinfection (1=entreprise ext., 0=éleveur)
	NETTOY	Type de main d'oeuvre pour le nettoyage (1=entreprise ext., 0=éleveur)
	TOTLIT	Quantité totale de litière utilisée pour le lot
	DENSI	Densité de poussins au début du lot (en $nombre/m^2$)
	SURF	Surface totale sur laquelle est élevé le lot (en m^2)
	VET	Montant des frais vétérinaires pour le lot (en $Euro/m^2$)
	GMQ	Indice de gain moyen quotidien
	ICE	Indice de consommation économique
	IP	Indice de performance
	KGM2	Poids moyen des dindes à l'abattage par unité de surface (en kg/m^2)
	ICT	Indice de consommation technique (en kg d'aliment/kg d'animal)

TAB. 5.1 – Description des variables relatives à l'étude des pertes économiques en élevage de dinde.

5.2 Description d'un tableau X orientée vers l'explication d'un tableau Y

5.2.1 Interprétation des composantes

Evolution des inerties expliquées par les composantes

La figure 5.1 illustre le pourcentage cumulé des variances des composantes $(t^{(1)}, \ldots, t^{(h)})$. Les résultats donnés par les méthodes *PCR* et *LRR* modifiée sont presque les mêmes ; les courbes apparaissent confondues sur cet exemple. Par la suite, pour simplifier l'interprétation, la version modifiée de la méthode *LRR* présentée paragraphe 3.1.2 page 51 est appelée *LRR*. La régression *PLS* donne des résultats comparables à ceux des méthodes *PCR* et *LRR*. Les variances des composantes de la régression *PLS* sont identiques à celles des méthodes *PCR* et *LRR* sur les trois premières composantes, et moins élevées pour les composantes suivantes. L'*ACPVI* a des composantes de variances nettement inférieures aux trois méthodes pré-citées.

La variance de la première composante $t^{(1)}$ peut être considérée comme une mesure de la stablité des modèles obtenus par la régression des variables Y sur la première composante. La figure 5.2 compare l'évolution de la variance de trois continuums en fonction de la valeur prise par leur paramètre (α ou γ_1), le continuum *LRR*, la méthode *principal covariate regression* et le continuum $ACPVI - PLS$. A titre

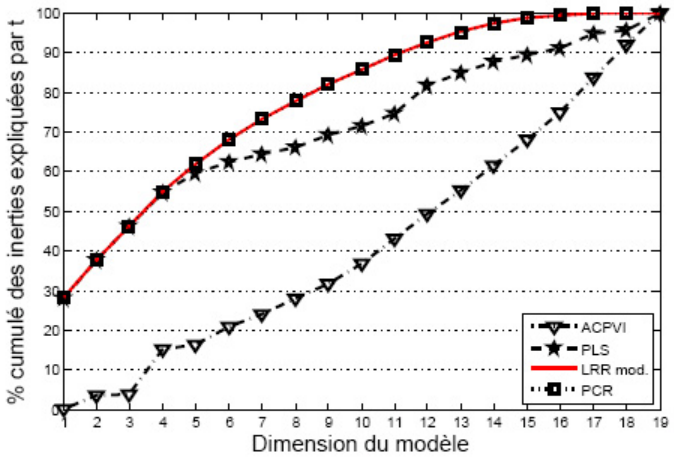


FIG. 5.1 – Pourcentage cumulé des inerties expliquées par les composantes $(t^{(1)},\ldots,t^{(h)})$. Comparaison des méthodes *ACPVI*, régression *PLS*, version modifiée de la *latent root regression* et régression sur composantes d'*ACP* (*PCR*).

illustratif, les cas particuliers connus de ces différents continuums, *PCR*, *LRR*, *PLS* et *ACPVI*, sont indiqués sur la figure 5.2. Pour le continuum $ACPVI-PLS$, nous retrouvons que $var(t^{(1)})$ est une fonction croissante du paramètre γ_1 (propriété démontrée paragraphe 4.2.2 page 70). La méthode *principal covariate regression* illustre le fait que la variance de la première composante d'*ACP* explique plus d'inertie que celle de l'*ACPVI*. Il faut noter que la contrainte de norme imposée par la méthode *principal covariate regression* est $\|t_\alpha\| = 1$; afin d'explorer et comparer la stabilité de ce continuum aux autres méthodes étudiées, nous avons modifié la standardisation de façon à avoir $\|w_\alpha\| = 1$. La variance de la première composante du continuum *LRR* est peu modifiée par la valeur prise par le paramètre α : les trois cas particuliers de ce continuum, *PCR*, *LRR* et *PLS*, définissent des composantes qui expliquent une inertie de l'ordre de 28%.

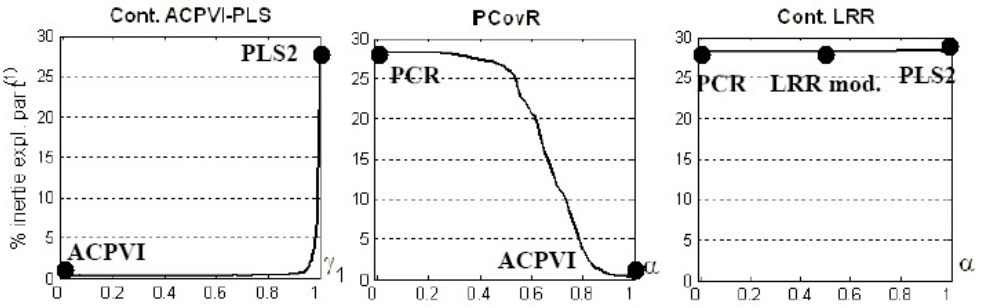


FIG. 5.2 – Evolution du pourcentage des inerties expliquées par la composante $t^{(1)}$ en fonction du paramètre du continuum (α ou γ_1) pour les méthodes : continuum *LRR*, *principal covariate regression* et continuum $ACPVI-PLS$. Les cas particuliers de ces continuums sont aussi indiqués.

Evolution des inerties des tableaux X et Y expliquées par les composantes

La figure 5.3 représente les parts des inerties des tableaux X et Y expliquées par les composantes. l'*ACPVI* explique légèrement moins bien le tableau X sur les trois premières dimensions. Les méthodes *PCR* et *LRR*, qui donnent des résultats comparables sur cet exemple, sont les méthodes dont les composantes expliquent le mieux les variables du tableau X. L'explication du tableau Y par ces mêmes composantes donne en revanche des résultats plus contrastés. L'*ACPVI* est la méthode qui explique le mieux les variables à expliquer (80% sur la première composante et 100% sur les deux premières composantes). Les méthodes *PCR* et *LRR* expliquent le moins le tableau Y et se différencient légèrement. La régression *PLS* donne des résultats intermédiaires entre ceux de l'*ACPVI* et la *PCR*, en termes d'explication de Y par les composantes. Ces résultats sont bien connus par les praticiens de l'analyse des données et confortent les observations faites sur les critères à maximiser relatifs à ces quatre méthodes (paragraphe 3.2.1 page 54).

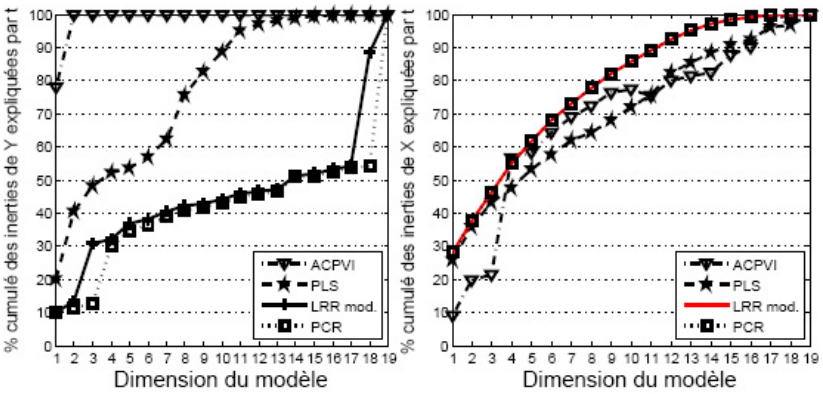


FIG. 5.3 – Pourcentage cumulé des inerties des tableaux Y et X expliquées par les composantes $(t^{(1)},\ldots,t^{(h)})$. Comparaison des méthodes *ACPVI*, régression *PLS*, version modifiée de la *latent root regression* et régression sur composantes de l'*ACP* (*PCR*).

L'inertie du tableau Y expliquée par la première composante $t^{(1)}$ peut être considérée comme une mesure de la qualité d'ajustement du modèle aux données. La figure 5.4 compare l'évolution de l'inertie de Y expliquée par $t^{(1)}$ des trois continuums étudiés, en fonction de leur paramètre (α ou γ_1). Pour le continuum $ACPVI-PLS$, nous retrouvons que cette inertie expliquée est une fonction décroissante du paramètre γ_1 (propriété démontrée paragraphe 4.2.2 page 70). Les méthodes dont le critère est plus orienté vers l'explication du tableau Y, c'est à dire l'*ACPVI* et dans une moindre mesure la régression *PLS*, apparaissent sur cet exemple comme ayant une meilleure qualité d'ajustement du modèle aux données. Le paramètre α du continuum *LRR* détermine le poids du tableau Y dans le critère à maximiser et semble avoir assez peu d'influence, dans cet exemple, sur la qualité d'ajustement de la méthode.

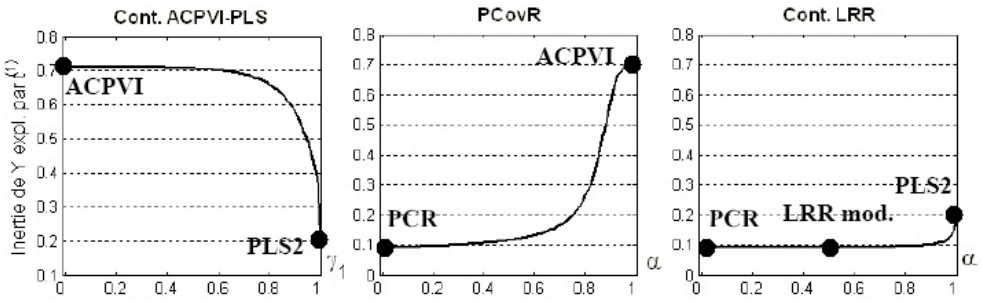
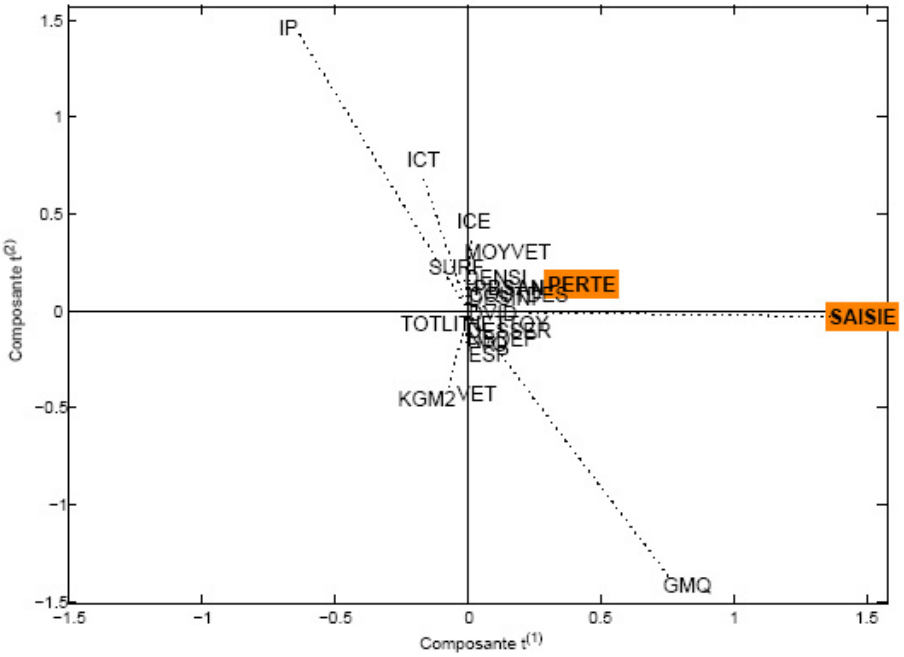


FIG. 5.4 – Comparaison des inerties de Y expliquées par la composante $t^{(1)}$ en fonction du paramètre du continuum (α ou γ_1) pour les méthodes : continuum *LRR*, *principal covariate regression* et continuum *ACPVI – PLS*. Les cas particuliers de ces continuums sont aussi indiqués.

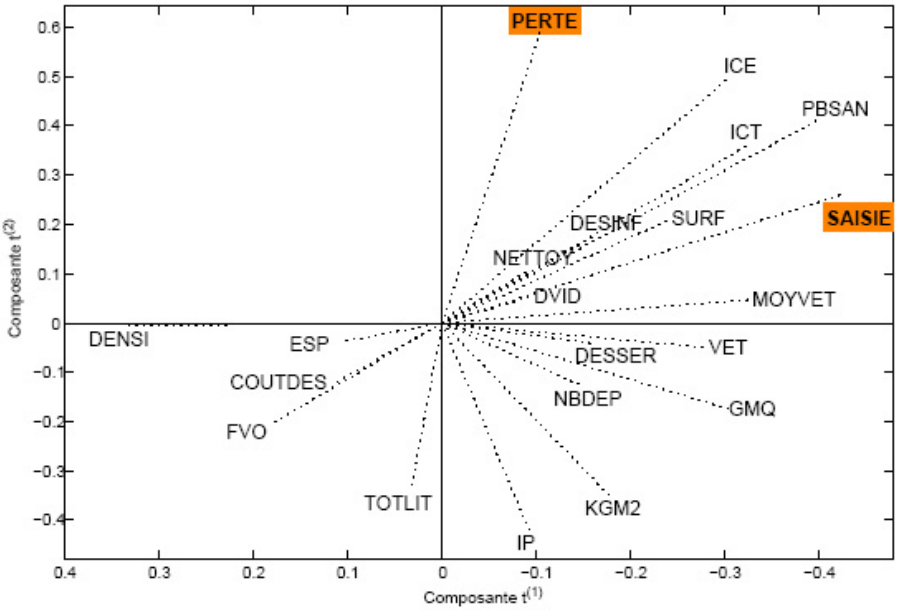
5.2.2 Représentation factorielle

Interprétation de la carte des variables

Du point de vue du praticien, les cartes factorielles consituent un outil précieux pour l'interprétation des liens entre variables, ainsi que pour l'explication des ressemblances et dissemblances entre les individus. Les variables de X et de Y peuvent être représentées sur le système formé par les composantes globales t orthogonales mutuellement. Les figures 5.5 et 5.6 illustrent ces représentations sur le plan des deux premières dimensions ($t^{(1)}$, $t^{(2)}$) pour les méthodes *ACPVI*, *PLS*, *LRR* et *ACP*. La carte de l'*ACPVI* diffère nettement des autres. Seule la variable à expliquer *SAISIE* semble être correctement expliquée sur ce plan factoriel. Les cartes factorielles des trois autres méthodes illustrent le fait que, sur ce plan, les deux variables Y apparaissent comme étant non corrélées et correctement expliquées par les variables X. La carte factorielle de l'*ACPVI* diffère aussi très nettement des autres cartes du point de vue de la représentation des variables X. Seules deux variables sont bien expliquées sur ce plan : l'indice de performance (*IP*) opposé au gain moyen quotidien (*GMQ*). Les cartes factorielles des autres méthodes, *ACP*, *latent root regression* et régression *PLS*, donnent des résultats dont l'interprétation est comparable sans être identique. Seule la carte de la régression *PLS* est interprétée par la suite. Afin de déterminer les facteurs de risque relatifs aux pertes économiques de l'éleveur, il est essentiel de raisonner conjointement sur les deux variables Y. Sur la carte factorielle de la régression *PLS*, les variables ayant des coordonnées négatives par rapport à la composante $t^{(1)}$ et positives par rapport à la composante $t^{(2)}$ sont associées à un profil à risque. Les indices relatifs aux consommations techniques et économiques *ICT* et *ICE*, ainsi que le fait qu'il y ait eu un problème sanitaire sérieux (*PBSAN*) sont des variables associées à de fortes pertes économiques pour l'éleveur. Les variables relatives à l'utilisation d'aliments exempts de farine de viande et d'os (*FVO*), au coût de la désinfection des bâtiments (*COUTDES*), et dans une moindre mesure à la quantité de litière utilisée pour le lot (*TOTLIT*) et au fait que le lot précédent soit de la même espèce (*ESP*), sont des facteurs associés à des pertes économiques plus faibles pour l'éleveur.

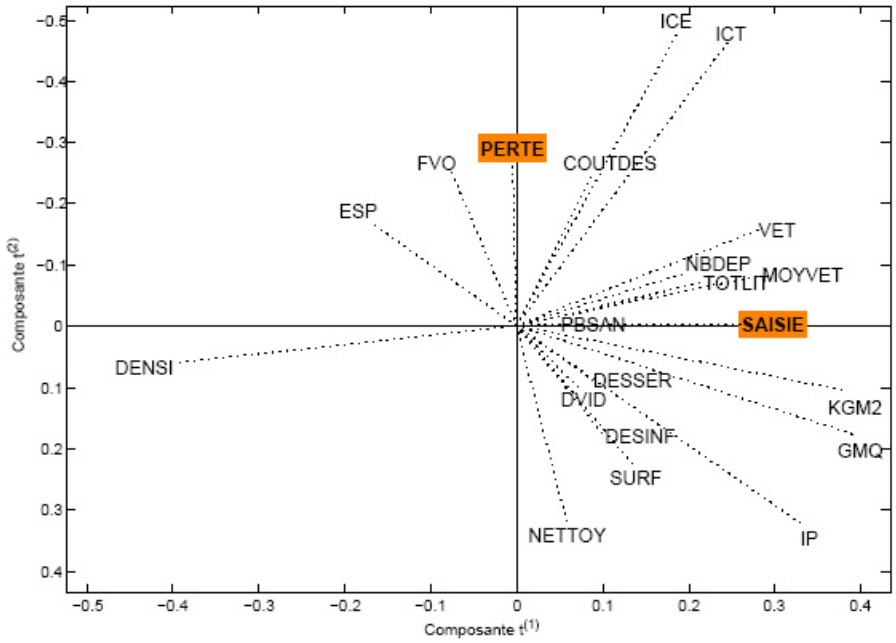


(a) *ACPVI*

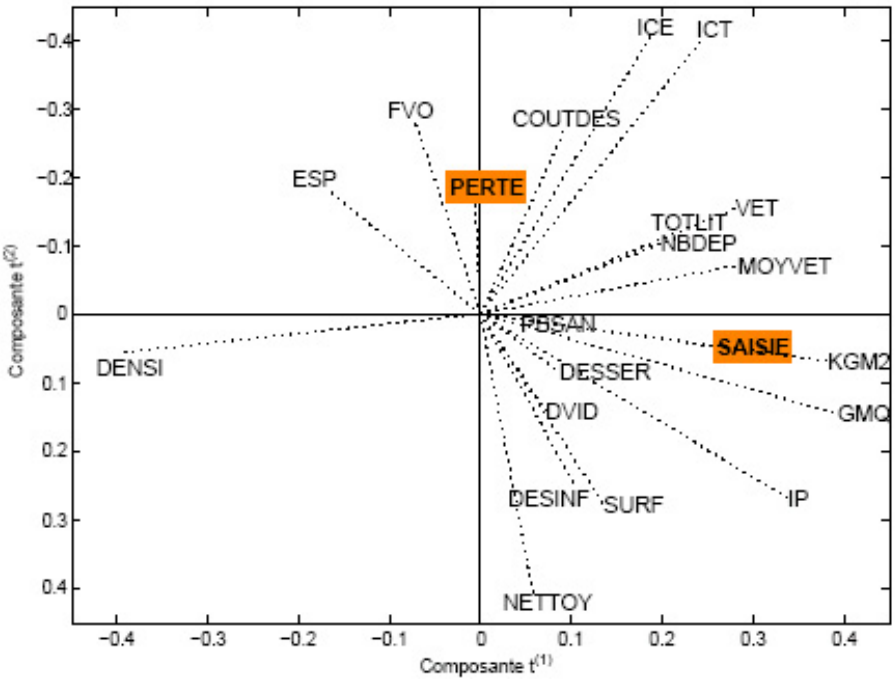


(b) Régression *PLS*

FIG. 5.5 – Représentation factorielle de l'ensemble des variables sur le plan des composantes $(t^{(1)}, t^{(2)})$.



(a) *LRR* modifiée



(b) *ACP*

FIG. 5.6 – Représentation factorielle de l'ensemble des variables sur le plan des composantes $(t^{(1)}, t^{(2)})$.

Interprétation du plan des individus

Il est intéressant de mettre en regard des cartes factorielles des variables (figures 5.5 et 5.6), les plans des individus qui leur sont associés, pour les quatre méthodes étudiées. Le plan des individus de l'*ACPVI* est présenté à partir des composantes normées. La couleur des individus sur la figure 5.7 est liée à la variable à expliquer *SAISIE*, qui représente le pourcentage de dindes saisies à l'abattoir. Les élevages colorés en jaune sont ceux pour lesquels la variables *SAISIE* a de faibles valeurs, et ceux qui sont colorés en rouge sont ceux qui ont des valeurs plus élevées pour cette même variable. Les plans factoriels des méthodes *PLS*, *LRR* et *ACP* présentent une répartition des individus assez peu orientée vers l'explication de la variable *SAISIE*. Les individus sont plutôt répartis, sur ce plan, selon trois groupes liés aux valeurs prises par les élevages pour les variables *IP*, *GMQ* et *KGM*2.

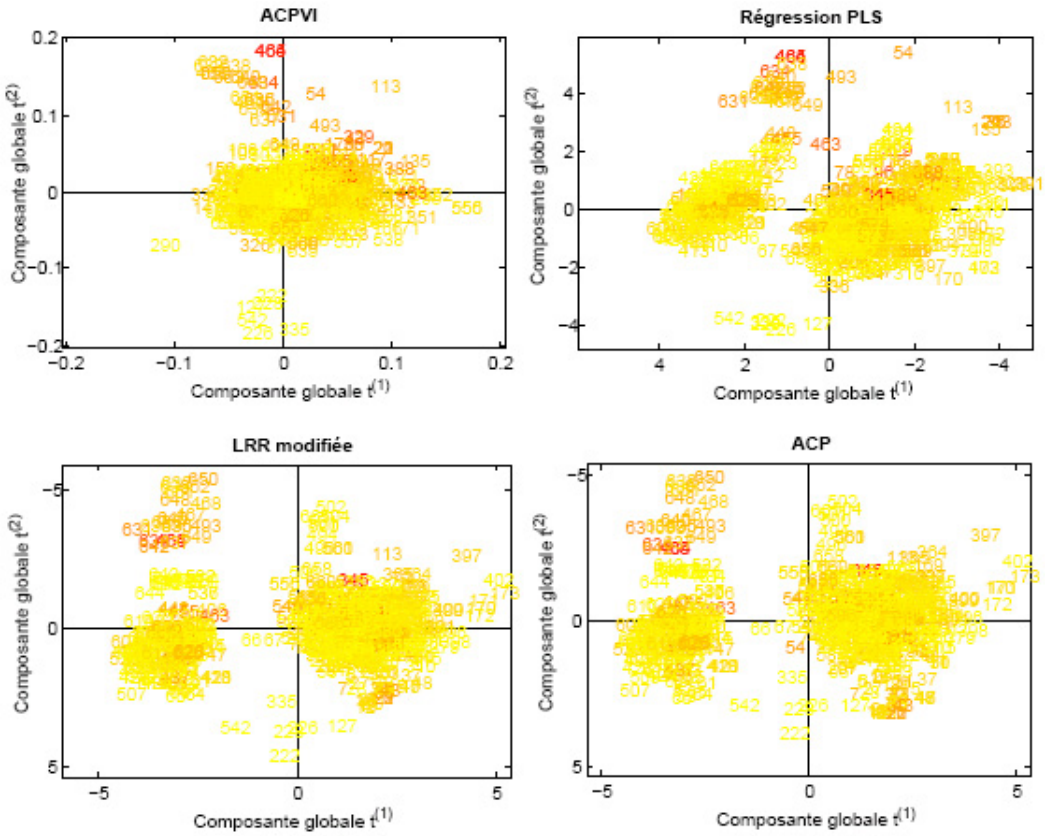


FIG. 5.7 – Représentation factorielle des individus sur le plan des composantes $(t^{(1)}, t^{(2)})$.

5.3 Prédiction de Y par X

5.3.1 Evolution de la norme du vecteur de coefficients

La norme du vecteur de coefficients de régression du modèle où une seule composante $t^{(1)}$ est retenue, $\|\beta^{(1)}\|$, varie en fonction des valeurs des paramètres des différents continuums (figure 5.8). Les continuums *ACPVI – PLS* et *principal covariate regression* donnent des valeurs de $\|\beta^{(1)}\|$ plus élevées pour les méthodes les plus orientées vers l'explication des variables Y ; il s'agit dans les deux cas de l'*ACPVI* comparativement soit à la régression *PLS* soit à la *PCR*. Le continuum *LRR*, sur cet exemple, présente peu de modification de la valeur de $\|\beta^{(1)}\|$ selon la valeur du paramètre α.

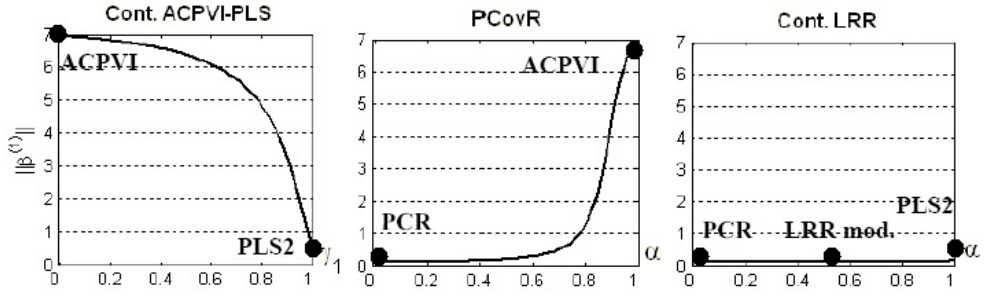


FIG. 5.8 – Evolution de la norme du vecteur de coefficients $\|\beta^{(1)}\|$ en fonction du paramètre du continuum (α ou γ_1) pour les méthodes : continuum *LRR*, *principal covariate regression* et continuum *ACPVI – PLS*. Les cas particuliers de ces continuums sont aussi indiqués.

5.3.2 Nombre optimal de dimensions

Résultats pour les méthodes de base

La figure 5.9 illustre l'évolution des erreurs moyennes de calibration et de validation selon le nombre de composantes conservées dans le modèle, pour les quatre méthodes étudiées. L'erreur moyenne de calibration ($RMSE_C$) illustre la qualité de l'ajustement du modèle aux données. Les résultats donnés par l'erreur de calibration sont comparables à ceux donnés par la figure 5.3, illustrant l'inertie du tableau Y expliquée par les composantes t. En effet, l'évolution de cette inertie est une autre façon de mesurer la qualité de l'ajustement du modèle aux données. L'erreur moyenne de validation ($RMSE_V$) illustre la qualité prédictive du modèle. Cette dernière est calculée grâce à la procédure de validation croisée décrite dans le paragraphe 3.2.2 page 55, sur la base de ($m = 500$) simulations. L'échantillon de calibration contient 2/3 des individus et l'échantillon de validation 1/3 des individus. Les résultats concernant la qualité d'ajustement et la qualité prédictive fournissent, sur cet exemple, les mêmes conclusions quant aux méthodes. Ceci est sûrement dû au grand nombre d'individus ($N = 659$) qui stabilise les résulats de la validation croisée. L'*ACPVI* est

la méthode ayant les meilleures performances sur cet exemple, malgré la multicolinéarité avérée des variables explicatives. Les méthodes *LRR* et *PCR* présentent, sur ce jeu de données, les moins bonnes performances et ne se différencient que sur les dernières dimensions. La régression *PLS* a des performances intermédiaires entre celles de l'*ACPVI* et de la *PCR*. Il faut noter que le modèle explicatif et prédictif optimal (minimisation des deux erreurs moyennes) est donné par l'*ACPVI* associée à deux composantes ($RMSE_C = 0.7289$ et $RMSE_V = 0.7751$).

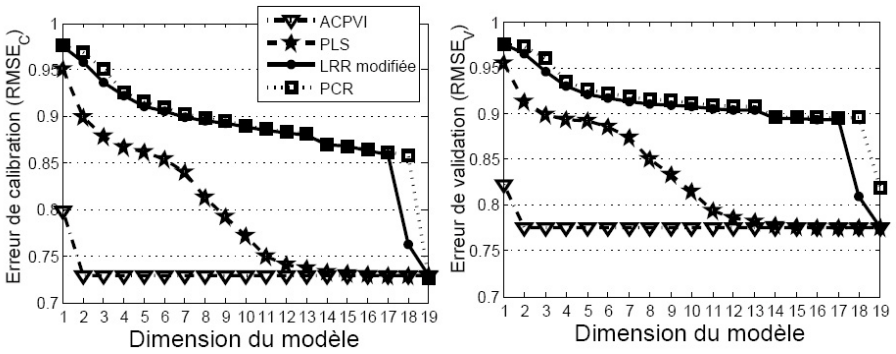


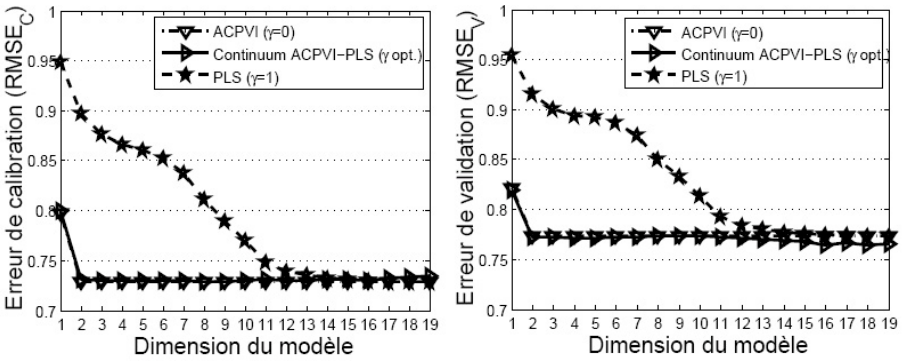
FIG. 5.9 – Erreur moyenne de calibration ($RMSE_C$) et de validation ($RMSE_V$) pour les méthodes *ACPVI*, régression *PLS*, *LRR* modifiée et *PCR*.

Résultats pour les approches continuums

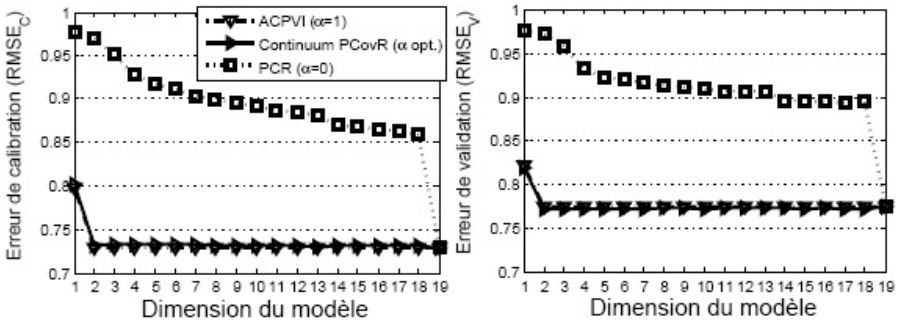
Les figures 5.10(a) à 5.10(c) illustrent l'évolution des erreurs moyennes de calibration et de validation selon le nombre de composantes conservées dans le modèle, pour les trois continuums étudiés. Ces erreurs sont calculées grâce à la même procédure validation croisée décrite dans le paragraphe 4.2.3 page 73, sur la base de ($m = 200$) simulations. Les paramètres α et γ_1 associés à chaque continuum varient entre 0 et 1 avec un pas de 0.01. Comme les paramètres des continuums sont déterminés de façon à minimiser l'erreur de validation, les continuums donnent des résultats optimaux, pour chaque dimension, pour la prédiction des variables Y par l'ensemble des variables X.

Les résultats du continuum $ACPVI - PLS$ et de ses cas particuliers sont décrits par la figure 5.10(a). Le continuum $ACPVI - PLS$ donne des résultats proches de ceux de l'*ACPVI*, méthode apparaissant optimale pour le traitement de ce jeu de données. Les valeurs prises par le paramètre γ_1 optimum sont de l'ordre de 0.2 pour les deux premières dimensions et à partir de la dimension 12. Pour les dimensions $h = (3, \ldots, 11)$, ces valeurs sont de l'ordre de 0.8, sans que les résultats du continuum ne se rapprochent de ceux de la régression *PLS*. Les résultats de la méthode *principal covariate regression* et de ses cas particuliers sont décrits par la figure 5.10(b). La méthode *principal covariate regression* donne aussi des résultats proches de ceux de l'*ACPVI*. Les valeurs du paramètre α optimum sont de l'ordre de 0.9 en moyenne, sauf pour les trois dernières dimensions où cette valeur diminue. Les résultats du continuum *LRR* et de ses cas particuliers sont décrits par la figure 5.10(c). Les erreurs moyennes de calibration et de validation de la *PCR* et de la version modifiée de la

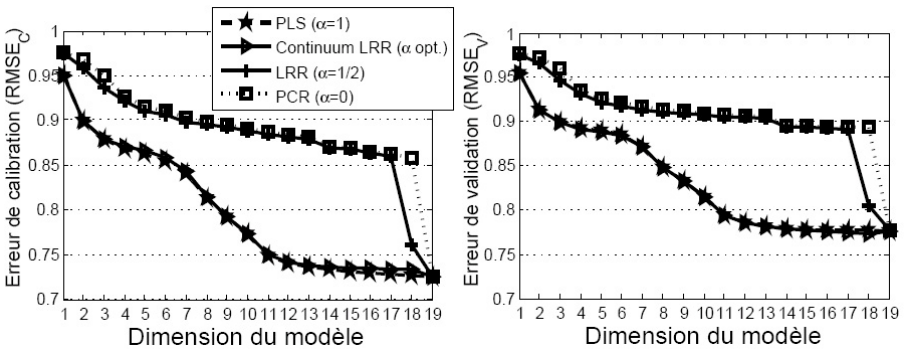
latent root regression sont proches sur cet exemple. Ces méthodes sont plus vouées à la description des variables X qu'à la prédiction de variables Y. Les résultats de la régression *PLS* sont plus performants en terme de prédiction des variables Y. Le continuum *LRR* donne des résultats proches de ceux de la régression *PLS*. Les valeurs moyennes du paramètre α optimum choisi par le continuum *LRR* sont de l'ordre de 0.9, sauf pour le modèle où toutes les dimensions sont retenues.



(a) Continuum *ACPVI – PLS*



(b) *Principal covariate regression*



(c) Continuum *LRR*

FIG. 5.10 – Erreurs moyennes de calibration ($RMSE_C$) et de validation ($RMSE_V$) obtenues pour chaque continuum et leurs cas particuliers.

Sélection du nombre optimal de dimensions pour chaque méthode de base

Les indices $Q^{2(h)}$ et $Q^{2(h)}_{cum}$ décrits dans le paragraphe 3.2.2 page 55 permettent de faire un compromis entre les résultats donnés par les erreurs moyennes de calibration et de validation. L'apport d'une composante $t^{(h)}$ dans le modèle peut être évalué par la valeur de l'indice $Q^{2(h)}$ (significatif au-delà du seuil empirique de 0.0975). L'apport des composantes $(t^{(1)},\ldots,t^{(h)})$ est évalué par la valeur de l'indice $Q^{2(h)}_{cum}$ (significatif au-delà du seuil empirique de 0.5). Les résultats, pour les quatre méthodes comparées, sont donnés dans le tableau 5.2. Les différences de résultat concernant la qualité d'ajustement et de prédiction apparaissent au travers des indices $Q^{2(h)}$ et $Q^{2(h)}_{cum}$. Seul l'apport de la première composante pour la méthode *ACPVI* est significatif.

Méthode	**Dimension** h	$Q^{2(h)}$	$Q^{2(h)}_{cum}$
ACPVI	1	0.3225 (*)	0.3225 (NS)
	2	0.0575 (NS)	0.3615 (NS)
	3	-0.131 (NS)	0.2778 (NS)
PLS	1	0.0882 (NS)	0.0882 (NS)
	2	0.0756 (NS)	0.1572 (NS)
	3	-0.0007 (NS)	0.1566 (NS)
LRR mod.	1	0.0458 (NS)	0.0458 (NS)
	2	0.0195 (NS)	0.0645 (NS)
	3	0.0244 (NS)	0.0873 (NS)
PCR	1	0.0452 (NS)	0.0452 (NS)
	2	0.0047 (NS)	0.0497 (NS)
	3	0.0165 (NS)	0.0654 (NS)

TAB. 5.2 – Valeurs des indices $Q^{2(h)}$ et $Q^{2(h)}_{cum}$ selon les trois premières dimensions du modèle pour les méthodes *ACPVI*, régression *PLS*, *LRR* modifiée et *PCR*. L'astérisque indique un apport significatif, *NS* indique un apport non significatif.

5.3.3 Poids des variables X dans l'explication de Y

Le poids des variables X est donné par les coefficients de régression de ces variables pour expliquer les variables Y. Afin de ne pas alourdir l'interprétation des résultats, seuls les coefficients de régression de la méthode *ACPVI* pour un modèle optimal à une dimension sont donnés. En effet, sur cet exemple, cette méthode fournit à la fois un bon ajustement et une bonne qualité de prédiction des données. 78% de la variabilité du tableau Y et 9% de la variabilité du tableau X sont expliquées par la composante $t^{(1)}$. Les écarts types et intervalles de variation à 95% associés aux coefficients de régression sont calculés à partir des résultats issus des ($m = 500$) simulations. On note que, du fait de la méthode de validation croisée choisie, ces indices de variabilité sont calculés sur 2/3 des individus.

Variables X	**SAISIE**	**PERTE**
FVO	**-0,06** [-0,10 ;-0,02]	-0,01 [-0,03 ;0,00]
COUTDES	0,00 [-0,04 ;0,04]	0,00 [-0,01 ;0,01]
MOYVET	0,06 [-0,02 ;0,14]	0,01 [-0,01 ;0,03]
DESSER	-0,01 [-0,04 ;0,02]	0,00 [-0,01 ;0,01]
DVID	0,02 [-0,02 ;0,06]	0,00 [0,00 ;0,01]
NBDEP	-0,01 [-0,05 ;0,03]	0,00 [-0,01 ;0,01]
ESP	0,02 [-0,02 ;0,06]	0,00 [0,00 ;0,01]
PBSAN	**0,16** [0,12 ;0,19]	**0,03** [0,02 ;0,05]
DESINF	0,01 [-0,03 ;0,04]	0,00 [-0,01 ;0,01]
NETTOY	-0,04 [-0,07 ;0,00]	-0,01 [-0,02 ;0,00]
TOTLIT	**-0,10** [-0,13 ;-0,06]	**-0,02** [-0,03 ;-0,01]
DENSI	-0,08 [-0,25 ;0,09]	-0,02 [-0,06 ;0,02]
SURF	0,04 [0,00 ;0,08]	0,01 [0,00 ;0,02]
VET	-0,02 [-0,11 ;0,07]	0,00 [-0,02 ;0,02]
GMQ	**5,10** [4,80 ;5,41]	**1,13** [0,72 ;1,55]
ICE	**0,07** [0,02 ;0,12]	0,02 [0,00 ;0,03]
IP	**-4,30** [-4,62 ;-3,98]	**-0,96** [-1,30 ;-0,61]
KGM2	**-0,51** [-0,64 ;-0,38]	**-0,11** [-0,17 ;-0,05]
ICT	**-1,14** [-1,35 ;-0,94]	**-0,25** [-0,35 ;-0,16]

TAB. 5.3 – Coefficients de régression et leurs intervalles de variation à 95%, du modèle liant X à Y, pour la méthode *ACPVI* avec une dimension.

L'interprétation de ces coefficients de régression est faite à l'aide des *odds ratio* et de leurs intervalles de variation (définis paragraphe 1.2.3 page 25), comme c'est l'usage pour les données d'épidémiologie animale. Un facteur de risque à effet protecteur est une variable explicative pour laquelle l'*odds ratio* est inférieur à un (valeur un non comprise dans l'intervalle de variation). A l'inverse, un facteur de risque est une variable explicative pour laquelle l'*odds ratio* est supérieur à un (valeur un non comprise dans l'intervalle de variation). Une variable explicative dont les *odds ratio* contiennent la valeur un dans leurs intervalles de variation n'est pas considérée comme ayant une influence significative sur les variables à expliquer.

Le sens et la significativité des *odds ratio* sont utilisés pour détecter les élévages ayant des profils à risque vis à vis des pertes économiques. Le profil d'un élevage où les pertes économiques sont élevées est celui où l'on a détecté :

1. un problème sanitaire sérieux durant la période d'élevage du lot de dindes,
2. un indice de gain moyen quotidien trop élevé,
3. un indice de consommation économique trop élevé, signe de problèmes sanitaires lors de l'élevage du lot (variable liée au taux de saisie à l'abattoir uniquement).

Le profil d'un élevage ayant de faibles pertes économiques est celui où l'on a détecté :

1. des aliments exempts de farines de viande et d'os (variable liée au taux de saisie à l'abattoir uniquement),
2. une quantité de litière limitée, dans le sens où l'utilisation d'une grande quantité de litière est un signe de problèmes digestifs lors de l'élevage du lot de

Variables X	**SAISIE**	**PERTE**
FVO	**0,94** [0,90 ;0,98]	0,99 [0,98 ;1,00]
COUTDES	1,00 [0,96 ;1,04]	1,00 [0,99 ;1,01]
MOYVET	1,06 [0,98 ;1,15]	1,01 [0,99 ;1,03]
DESSER	0,99 [0,96 ;1,02]	1,00 [0,99 ;1,01]
DVID	1,02 [0,98 ;1,06]	1,00 [1,00 ;1,01]
NBDEP	0,99 [0,95 ;1,03]	1,00 [0,99 ;1,01]
ESP	1,02 [0,98 ;1,06]	1,00 [1,00 ;1,01]
PBSAN	**1,17** [1,13 ;1,21]	**1,04** [1,02 ;1,05]
DESINF	1,01 [0,97 ;1,04]	1,00 [0,99 ;1,01]
NETTOY	0,96 [0,93 ;1,00]	0,99 [0,98 ;1,00]
TOTLIT	**0,91** [0,88 ;0,94]	**0,98** [0,97 ;0,99]
DENSI	0,92 [0,78 ;1,10]	0,98 [0,95 ;1,02]
SURF	1,04 [1,00 ;1,08]	1,01 [1,00 ;1,02]
VET	0,98 [0,90 ;1,07]	1,00 [0,98 ;1,02]
GMQ	**164,35** [121,24 ;222,78]	**3,11** [2,05 ;4,72]
ICE	**1,07** [1,02 ;1,13]	1,02 [1,00 ;1,03]
IP	**0,01** [0,01 ;0,02]	**0,38** [0,27 ;0,54]
KGM2	**0,60** [0,53 ;0,69]	**0,89** [0,84 ;0,95]
ICT	**0,32** [0,26 ;0,39]	**0,78** [0,71 ;0,85]

TAB. 5.4 – *Odds ratio* et leurs intervalles de variation à 95%, du modèle liant X à Y, pour la méthode *ACPVI* avec une dimension. Les *OR* significatifs sont en gras.

dindes,

3. un indice de performance plutôt faible,
4. un poids moyen de dindes à l'abattage rapporté à la surface disponible au sol (appelé aussi chargement) plutôt faible,
5. un indice de consommation technique plutôt faible. Cet indice représente la quantité d'aliment consommé rapporté au poids final de l'animal. Une faible valeur de cet indice indique qu'il n'y a pas de problème sanitaire de type digestif et que les dindes ont consommé une quantité raisonnable d'aliment pour atteindre leur poids final.

Troisième partie

Description de K tableaux X_k orientée vers l'explication d'un tableau Y

Chapitre 6

Analyse de $(K+1)$ tableaux

6.1 Méthodes liant K tableaux X_k à un tableau Y

6.1.1 Format des données et objectifs

Les études d'épidémiologie analytique dans le domaine animal sont structurées en groupes (paragraphe 2.2.2 page 39) : les caractéristiques de la ferme (taille de l'élevage, performances zootechniques, autres productions animales, ...), la conduite d'élevage (taux de renouvellement, technique de reproduction, nombre d'animaux par portée, nombre de bandes d'animaux, ...), l'habitat des animaux (enregistrements bio-climatiques, ventilation, isolation, chauffage, ...), l'alimentation et l'abreuvement des animaux (compositions alimentaires, mode de distribution, nombre de mangeoires, origine des aliments, ...), l'état sanitaire du troupeau (dosages sérologiques, pesées, maladies chroniques, taux de réformes, vaccinations, traitements antibiotiques, ...), les pratiques d'hygiène (protocoles de nettoyage et de désinfection) et les mesures de bio-sécurité (mesures sanitaires de l'éleveur et des visiteurs, équarrissage, ...). Toutes ces variables sont potentiellement explicatives d'une maladie caractérisée par plusieurs variables.

Les objectifs du traitement statistique des données d'épidémiologie animale sont à la fois descriptifs et prédictifs. Des représentations factorielles synthétiques, orientées vers l'explication de la maladie permettent de comprendre les liens complexes entre variables. L'explication de la maladie Y par les variables explicatives X est réalisée en parallèle à l'étude descriptive et permet de déterminer les variables qui réduisent l'apparition ou la prévalence de la maladie. Du fait du grand nombre de variables X et du sens biologique des blocs dans lesquels celles-ci sont organisées, il apparaît utile de mesurer à la fois l'importance des variables mais aussi des blocs de variables dans l'explication de la maladie. Le développement de méthodes statistiques en vue du traitement des données d'épidémiologie animale doit prendre en compte à la fois la particularité des données et des objectifs de traitement associés. Pour l'ensemble des méthodes présentées par la suite, les objectifs sont comparables : la problématique statistique relève à la fois de la description et de la modélisation à partir de tableaux multiples, soit K tableaux X_k orientés vers l'explication d'un tableau Y. Il s'agit tout d'abord de décrire simultanément un ensemble de K tableaux X_k décrivant les mêmes individus, ce qui revient à déterminer ce qui est commun à

l'ensemble ou à une partie des tableaux dans le but de décrire Y et, par conséquent, ce qui différencie certains tableaux des autres.

Nous disposons d'un tableau $X = [x_1, \ldots, x_P]$ contenant P variables explicatives et d'un tableau $Y = [y_1, \ldots, y_Q]$ contenant Q variables à expliquer. Ces deux tableaux sont mesurés sur les mêmes N individus. Le tableau X est partitionné en K blocs : $X = [X_1|\ldots|X_K]$. Chaque tableau X_k est un tableau $(N \times p_k)$ dont les lignes correspondent aux mêmes individus. La structure des données en $(K+1)$ tableaux est illustrée par la figure 6.1.

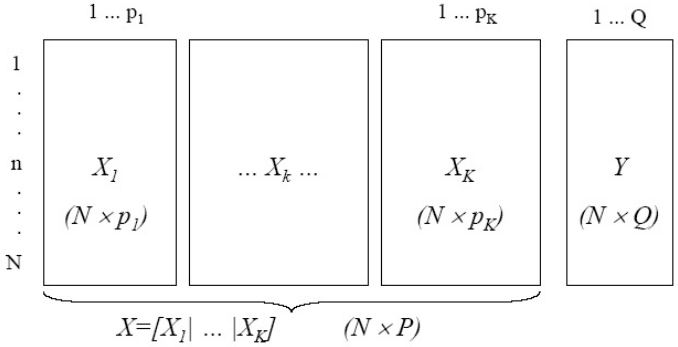


FIG. 6.1 – Illustration de la structure des $(K+1)$ tableaux X_k pour $k = (1, \ldots, K)$ et Y.

L'ensemble des méthodes décrites par la suite est basé sur l'extraction de composantes globales résumant le tableau concaténé $X = [X_1|\ldots|X_K]$, synthèses de composantes partielles résumant elles-mêmes chacun des tableaux $(X_1, \ldots, X_K)$. Les composantes globales résumant X sont orientées vers l'explication de Y. Le lien entre les tableaux et leurs composantes associées est illustré par la figure 6.2. Le tableau concaténé X joue le rôle du tableau compromis qui exprime la structure commune et spécifique des K tableaux X_k pour expliquer Y. Le tableau X fournit des composantes globales sur lesquelles se projettent l'ensemble des variables des tableaux X_k et Y. Pour cela, différentes approches sont proposées privilégiant soit le cadre de l'analyse canonique (travaux de Kissita [2003]), soit de nouvelles voies issues de la généralisation de l'*ACPVI*, ou des méthodes moins vulnérables à la multicolinéarité comme des solutions issues de la régression *PLS* (travaux de Vivien [2002]) ou de la généralisation de la *latent root regression*.

6.1.2 Méthodes s'apparentant à l'analyse canonique

Analyse canonique généralisée avec un tableau de référence

L'analyse canonique (paragraphe 3.1.3 page 52) est généralisée pour le cas de K tableaux $(X_1, \ldots, X_K)$ par Horst [1961]. Carroll [1968] modifie le critère en introduisant une composante globale (appelée variable auxiliaire), résumé du tableau concaténé $X = [X_1|\ldots|X_K]$, ce qui aboutit à une solution non itérative de la maximisation du critère. Des variantes de cette solution sont proposées par d'autres auteurs [Kettenring, 1971; Saporta, 1975; Meyer, 1989; Pontier et Normand, 1992; Casin, 1995].

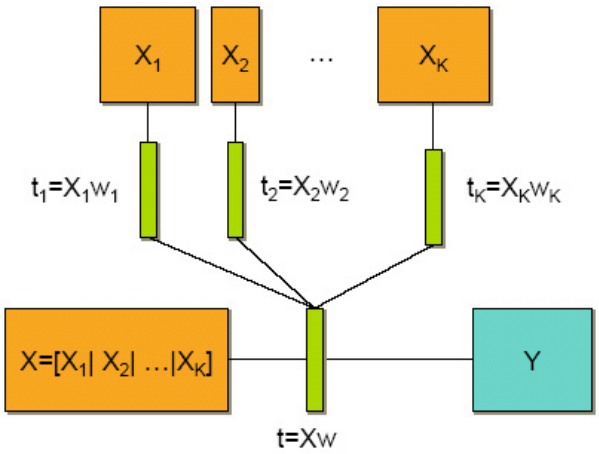


FIG. 6.2 – Illustration des liens entre K tableaux X_k ($k = 1,\ldots,K$) et un tableau Y, résumés chacun par une composante, pour une dimension donnée.

Une synthèse est proposée par Hanafi [1997] et Hanafi et Kiers [2006]. Cette généralisation, appelée analyse canonique généralisée ou *ACG* [Carroll, 1968], recherche les composantes partielles t_k, résumés de chaque tableau X_k, les plus liées à une composante globale t, au sens de la corrélation. Elle est basée sur la maximisation du critère (6.1).

$$\sum_{k=1}^{K} cov^2(t_k^{(1)}, t^{(1)}) \quad \text{avec} \quad t_k^{(1)} = X_k w_k^{(1)},\ t^{(1)} = Xw^{(1)},\ \|t_k^{(1)}\| = \|t^{(1)}\| = 1 \tag{6.1}$$

Kissita [2003, Chap. 3] propose une extension de l'analyse canonique généralisée dont l'objectif est de lier K tableaux $(X_1,\ldots,X_K)$ à un tableau Y. Cette méthode est appelée analyse canonique généralisée avec tableau de référence (*ACGTR*). L'*ACGTR* recherche, dimension par dimension, les composantes t_k, résumés de chaque tableau X_k, et u, résumé du tableau Y, dont la somme des carrés des corrélations est maximum. Cette méthode permet d'étudier les dépendances simultanées entre les K tableaux X_k et le tableau de référence Y. Elle est basée sur la maximisation du critère (6.2). Ce critère s'apparente à une analyse canonique des tableaux X_k, où la variable auxiliaire est contrainte d'être dans l'espace des variables de Y.

$$\sum_{k=1}^{K} cov^2(t_k^{(1)}, u^{(1)}) \quad \text{avec} \quad t_k^{(1)} = X_k w_k^{(1)},\ u^{(1)} = Yv^{(1)},\ \|t_k^{(1)}\| = \|u^{(1)}\| = 1 \tag{6.2}$$

En s'inspirant du problème de la double optimisation de Lafosse et Hanafi [1997, Prop. 3.1] développé dans le cadre de la généralisation de l'analyse factorielle interbatterie, Kissita [2003] démontre que la maximisation du critère (6.2) est équivalente à celle du critère (6.3) qui fait apparaître une composante globale t liée au tableau concaténé $X = [X_1|\ldots|X_K]$:

$$cov^2(t^{(1)}, u^{(1)}) \quad \text{avec} \quad u^{(1)} = Yv^{(1)}, \quad t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)},\ t_k^{(1)} = X_k w_k^{(1)} \tag{6.3}$$

$$\textstyle\sum_k a_k^{(1)2} = 1, \quad t^{(1)} = Xw^{(1)}, \quad \|t_k^{(1)}\| = \|u^{(1)}\| = 1$$

La solution de ce problème de maximisation est donnée par $w^{(1)}$ vecteur propre de la matrice $(1/N^2)(X'X)^{-1}X'Y(Y'Y)^{-1}Y'X$ associé à la plus grande valeur propre $\lambda^{(1)}$ [Kissita, 2003, p. 50]. Par la suite, $w^{(1)}$ est partitionné en blocs conformément à la partition de X en K tableaux : $w^{(1)} = [w_1^{(1)'}|\ldots|w_K^{(1)'}]'$. A partir de là, les composantes $t_k^{(1)}$ normées sont données à partir de $t_k^{*(1)} = X_k w_k^{(1)}/\|X_k w_k^{(1)}\|$. La composante $u^{(1)}$ est calculée à partir de $u^{(1)} = Yv^{(1)}$, avec $v^{(1)}$ vecteur propre de la matrice $(1/N^2)(Y'Y)^{-1}Y'X(X'X)^{-1}X'Y$ associé à la plus grande valeur propre $\lambda^{(1)}$. Les solutions d'ordre suivant sont obtenues par déflation des tableaux X_k sur leurs composantes t_k respectives. Les composantes partielles $(t_k^{(1)},\ldots,t_k^{(h)})$, pour chaque valeur de k, sont donc mutuellement orthogonales par construction. Kissita [2003, p. 129] précise que les composantes $(u^{(1)},\ldots,u^{(h)})$ sont aussi mutuellement orthogonales, mais que les composantes globales $(t^{(1)},\ldots,t^{(h)})$ ne le sont pas.

Il ressort des solutions de l'*ACGTR* que les composantes globales $t^{(1)}$ et $u^{(1)}$, associées respectivement au tableau concaténé X et au tableau Y, ne dépendent pas de la partition de X en K blocs. L'*ACGTR* apporte une solution théorique intéressante au lien entre $(K+1)$ tableaux, mais présente les mêmes limites pratiques que l'analyse canonique en cas de multicolinéarité au sein du tableau concaténé X ou au sein du tableau Y (paragraphe 3.1.3 page 52 ou Lebart *et al.* [2000, p. 352]).

Analyse canonique généralisée sous contrainte

Comme nous l'avons indiqué dans le paragraphe précédent, l'analyse canonique généralisée est basée sur la maximisation du critère (6.1). Nous proposons une modification de ce critère qui consiste à maximiser le même critère en modifiant les contraintes qui lui sont associées, selon le problème de maximisation (6.4).

$$\sum_{k=1}^{K} cov^2(t_k^{(1)}, t^{(1)}) \quad \text{avec} \quad t_k^{(1)} = X_k w_k^{(1)},\ t^{(1)} = Xw^{(1)},\ \|t_k^{(1)}\| = \|w^{(1)}\| = 1 \tag{6.4}$$

Pour une composante $t^{(1)}$ fixée, la valeur optimale de la composante partielle $t_k^{(1)}$ est donnée par $t_k^{(1)} = P_k t^{(1)}/\|P_k t^{(1)}\|$, avec $P_k = X_k(X_k'X_k)^{-1}X_k'$. Les composantes partielles $t_k^{(1)}$ sont les composantes normées issues de la projection de la composante $t^{(1)}$ sur les espaces associés aux tableaux X_k pour $k = (1,\ldots,K)$. En reportant cette valeur dans l'expression (6.4), le critère devient $\sum_k cov^2(t_k^{(1)}, t^{(1)}) = t^{(1)'}P_k t^{(1)} = \sum_k w^{(1)'}X'P_k X w^{(1)}$. Ainsi, $w^{(1)}$ est le premier vecteur propre de la matrice $H = (1/N^2)X'[\sum_k X_k(X_k'X_k)^{-1}X_k']X$ associé à la plus grande valeur propre. Les composantes d'ordre suivant s'obtiennent après déflation sur les composantes t obtenues aux étapes précédentes. Etant donné que les projecteurs P_k sont symétriques et idempotents, il s'ensuit que $H = (1/N^2)\sum_k (P_k X)'(P_k X)$. Cette version modifiée de l'*ACG* consiste donc à réaliser une *ACP* du tableau obtenu par concaténation verticale des projections de X sur les espaces engendrés par les blocs X_k. Nous pouvons remarquer que cette version de l'analyse canonique généralisée permet de limiter les problèmes de sensibilité à la multicolinéarité des variables du tableau concaténé X car la matrice $(X'X)$ n'a pas besoin d'être inversée.

A partir de cette version modifiée (6.4) du critère de l'*ACG*, nous définissons une analyse canonique généralisée sous contrainte. Cette méthode a les mêmes objectifs que l'*ACGTR*, c'est à dire, déterminer simultanément des composantes globales t résumant le tableau concaténé $X = [X_1|\ldots|X_K]$ et orientées vers l'explication du tableau Y, ainsi que des composantes partielles t_k résumant respectivement chaque tableau X_k et liées au tableau concaténé X. La méthode proposée est une solution intermédiaire entre celle de l'*ACG* adaptée à la description de K tableaux X_k, et celle de l'*ACGTR* qui permet la description de K tableaux X_k orientée vers l'explication d'un tableau Y, en appliquant la contrainte de norme sur l'axe w et non plus sur la composante t. La solution de cette méthode est donnée par la maximisation du critère (6.5), où le tableau Y est résumé par une composante $u^{(1)}$.

$$cov^2(u^{(1)},t^{(1)}) + \sum_k cov^2(t_k^{(1)},t^{(1)}) \quad \text{avec} \quad t_k^{(1)} = X_k w_k^{(1)} \tag{6.5}$$
$$t^{(1)} = Xw^{(1)}, \quad u^{(1)} = Yv^{(1)}, \quad \|t_k^{(1)}\| = \|w^{(1)}\| = \|u^{(1)}\| = 1$$

Pour $t^{(1)}$ fixée, la valeur optimale de $u^{(1)}$ est $u^{(1)} = Y(Y'Y)^{-1}Y't^{(1)}/\|Y(Y'Y)^{-1}Y't^{(1)}\|$, composante normée issue de la projection de la composante $t^{(1)}$ sur l'espace associé au tableau Y. Les composantes normées $t_k^{(1)}$ sont obtenues par projection de la composante $t^{(1)}$ sur les espaces associés aux tableaux X_k : $t_k^{(1)} = X_k(X_k'X_k)^{-1}X_k't^{(1)}/\|X_k(X_k'X_k)^{-1}X_k't^{(1)}\|$ pour $k = (1,\ldots,K)$. En reportant ces valeurs dans l'expression (6.5), il s'ensuit que :

$$\begin{aligned}
(6.5) &= (1/N^2)\left[t^{(1)'}u^{(1)}u^{(1)'}t^{(1)} + \sum_k t^{(1)'}t_k^{(1)}t_k^{(1)'}t^{(1)}\right] \\
&= (1/N^2)\left[w^{(1)'}X'Y(Y'Y)^{-1}Y'Xw^{(1)} + w^{(1)}X'(\sum_k X_k(X_k'X_k)^{-1}X_k')Xw^{(1)}\right] \\
&= (1/N^2)w^{(1)'}\left[X'Y(Y'Y)^{-1}Y'X + X'[\sum_k X_k(X_k'X_k)^{-1}X_k']X\right]w^{(1)}
\end{aligned}$$

La solution de cette maximisation est donnée par $w^{(1)}$, premier vecteur propre de la matrice $(1/N^2)\left[X'\left(Y(Y'Y)^{-1}Y' + \sum_k X_k(X_k'X_k)^{-1}X_k'\right)X\right]$ associé à la plus grande valeur propre $\lambda^{(1)}$. Les axes et composantes d'ordre suivant sont issus de la maximisation du critère (6.5) en considérant les résidus successifs de la régression des tableaux X_k et Y sur les composantes globales t obtenues aux étapes précédentes. Le modèle expliquant le tableau Y par l'ensemble des variables du tableau concaténé $X = [X_1|\ldots|X_K]$ s'appuie sur les composantes $(t^{(1)},\ldots,t^{(h)})$, qui sont mutuellement orthogonales par construction.

6.1.3 Extensions de l'*ACPVI* au cas de $(K+1)$ tableaux

Dans le cadre des méthodes s'apparentant à l'analyse canonique présentées dans le paragraphe 6.1.2, il est clair que l'accent est mis sur l'investigation des relations entre Y et les tableaux X_k, sans se soucier de la capacité des composantes à restituer l'inertie du tableau Y. Dans ce paragraphe, nous focalisons toujours sur

l'analyse des relations entre les tableaux Y et X_k, en mettant cette fois-ci l'accent sur la restitution de l'inertie de Y. Ces méthodes apparaissent donc davantage appropriées pour répondre aux objectifs poursuivis en épidémiologie animale (paragraphe 2.3 page 41). Nous proposons des méthodes, dérivant d'extensions de l'Analyse en Composantes Principales sur Variables Instrumentales ou *ACPVI* (paragraphe 3.1.1 page 47), dans le cas où le tableau X est structuré en K blocs. Des extensions de l'*ACPVI* pour plus de deux tableaux sont évoquées, mais non détaillées, par Hanafi et Lafosse [2001] ainsi que Kissita *et al.* [2004, p. 82] dans les travaux relatifs à l'analyse de concordance généralisée avec une métrique de Mahalanobis.

ACPVI **des tableaux** Y **et** X_k

Afin de déterminer les composantes partielles t_k, synthèses des liens entre chaque tableau X_k et le tableau Y, une première solution consiste à maximiser le critère (6.6).

$$\sum_{k=1}^{K}\sum_{q=1}^{Q} cov^2(y_q, t_k^{(1)}) \quad \text{avec} \quad t_k^{(1)} = X_k w_k^{(1)} \quad \text{et} \quad \|t_k^{(1)}\| = 1 \tag{6.6}$$

La solution de ce problème revient en définitive à effectuer de manière indépendante K *ACPVI* des couples de tableaux (Y, X_k) pour $k = (1, \ldots, K)$. Cette méthode ne répond qu'à une partie des objectifs formulés. Les composantes partielles obtenues sont intéressantes du point de vue de l'explication des liens entre chaque tableau X_k et le tableau Y. Cependant, ces *ACPVI* restituent l'inertie de Y dans des directions propres à chaque tableau X_k. Les différentes composantes étant déterminées de manière indépendante les unes des autres ne véhiculent pas nécessairement des informations concordantes. De ce fait, l'interprétation des résultats peut s'avérer fastidieuse par manque d'une vision synthétique.

ACPVI **des tableaux** Y **et** $X = [X_1|\ldots|X_K]$

Une seconde solution consiste à chercher à la fois les composantes globales $t^{(1)} = Xw^{(1)}$ associées au tableau concaténé $X = [X_1|\ldots|X_K]$ et les composantes partielles $t_k^{(1)} = X_k w_k^{(1)}$ associées aux tableaux X_k, de manière à maximiser le critère (6.7).

$$\sum_{q=1}^{Q} cov^2(y_q, t^{(1)}) \quad \text{avec} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}, \quad t_k^{(1)} = X_k w_k^{(1)}, \quad \|t^{(1)}\| = \|t_k^{(1)}\| = 1 \tag{6.7}$$

La solution de ce problème est donnée par l'*ACPVI* de Y par rapport au tableau concaténé X. En effet, soit $t^{(1)} = Xw^{(1)}$ avec $w^{(1)}$, vecteur propre associé à la plus grande valeur propre de la matrice $(1/N^2)[(X'X)^{-1}X'YY'X]$. Ce vecteur $w^{(1)}$ est partitionné en K sous-vecteurs $w_k^{(1)}$ relatifs aux différents tableaux X_k de la façon suivante : $w^{(1)} = [w_1^{(1)}|\ldots|w_K^{(1)}]'$. Si l'on pose $t_k^{(1)*} = X_k w_k^{(1)}$, la solution du problème (6.7) consiste à prendre $t_k^{(1)} = t_k^{(1)*}/\|t_k^{(1)*}\|$ et $a_k^{(1)} = \|t_k^{(1)*}\|$. Cette solution, à première vue satisfaisante, présente toutefois un inconvénient. La méthode étant basée sur

l'*ACPVI* des tableaux Y et X, elle ne prend pas en compte la structure en blocs du tableau X pour la détermination des composantes globales t. De plus, elle implique l'inversion de la matrice de variance-covariance $(X'X)$. Ceci pourrait constituer un problème lié à l'instabilité des résultats lorsque les variables du tableau concaténé X présentent des quasi-colinéarités. Le problème se pose de manière plus aiguë que précédemment, car même s'il n'y a pas a priori de redondance à l'intérieur des blocs X_k pris séparément, il se peut que la concaténation des tableaux X_k engendre des problèmes de quasi-colinéarité. Afin de contourner cette difficulté souvent rencontrée en pratique, nous proposons une démarche basée sur un critère original.

ACPVI **multibloc**

Comme dans le paragraphe précédent, la méthode proposée permet d'extraire directement une composante globale $t^{(1)} = Xw^{(1)}$ orientée vers l'explication du tableau Y, cette composante étant contrainte à être la synthèse des composantes partielles $t_k^{(1)}$. Le critère (6.8) ci-après constitue une variante du critère (6.7), car il est basé sur la même fonction à maximiser, mais d'autres contraintes sont imposées. Le problème (6.8) consiste à maximiser :

$$\sum_{q=1}^{Q} cov^2(y_q, t^{(1)}) \quad \text{avec} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}, \quad t_k^{(1)} = X_k w_k^{(1)}, \tag{6.8}$$
$$\sum_{k=1}^{K} a_k^{(1)2} = 1 \quad \text{et} \quad \|t_k^{(1)}\| = 1$$

Afin de déterminer les solutions de ce problème, nous proposons de développer des propriétés liées au critère (6.8). Dans un premier temps, nous montrons que la composante t, solution du critère (6.8), est également solution du critère (6.9) :

$$cov^2(u^{(1)}, t^{(1)}) \quad \text{avec} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}, \quad t_k^{(1)} = X_k w_k^{(1)}, \tag{6.9}$$
$$u^{(1)} = Yv^{(1)}, \quad \sum_{k=1}^{K} a_k^{(1)2} = 1 \quad \text{et} \quad \|t_k^{(1)}\| = \|v^{(1)}\| = 1$$

L'avantage de cette formulation est d'exhiber une composante $u = Yv$ dans l'espace engendré par les variables de Y, liée de manière optimale à la composante globale t. Pour la démonstration de cette propriété, nous pouvons remarquer que la maximisation de $cov^2(u^{(1)}, t^{(1)}) = cov^2(Yv^{(1)}, t^{(1)})$ par rapport à $v^{(1)}$, conduit à la solution $v^{(1)} = Y't^{(1)}/\|Y't^{(1)}\|$. En remplaçant cette solution dans l'expression $cov^2(Yv^{(1)}, t^{(1)})$, nous trouvons $cov^2(u^{(1)}, t^{(1)}) = (1/N^2)t^{(1)'}YY't^{(1)} = \sum_q cov^2(y_q, t^{(1)})$. Le critère (6.10) ci-après, permet de montrer que la composante $u^{(1)}$ est liée aux composantes partielles $t_k^{(1)}$ pour $k = (1, \ldots, K)$. Ce critère stipule que les composantes optimales $t_k^{(1)}$ des problèmes (6.8) et (6.9), sont également solutions du problème (6.10).

$$\sum_{k=1}^{K} cov^2(u^{(1)}, t_k^{(1)}) \quad \text{avec} \quad t_k^{(1)} = X_k w_k^{(1)}, \quad u^{(1)} = Yv^{(1)}, \tag{6.10}$$
$$\|t_k^{(1)}\| = \|v^{(1)}\| = 1$$

En effet, en remplaçant dans le critère (6.9) la composante $t^{(1)}$ par $\sum_k a_k^{(1)} t_k^{(1)}$, nous obtenons $cov^2(t^{(1)}, u^{(1)}) = \left[\sum_k a_k^{(1)} cov(u^{(1)}, t_k^{(1)})\right]^2$. La maximisation de ce critère par rapport à $a_k^{(1)}$ pour $k = (1, \ldots, K)$, conduit à $a_k^{(1)} = cov(u^{(1)}, t_k^{(1)}) / \sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})}$. En remplaçant cette valeur dans l'expression ci-dessus, nous obtenons : $cov^2(t^{(1)}, u^{(1)}) = \sum_k cov^2(u^{(1)}, t_k^{(1)})$. C'est à partir du critère (6.10) que la solution du problème est développée. Nous avons :

$$\begin{aligned} \sum_k cov^2(u^{(1)}, t_k^{(1)}) &= (1/N^2) \sum_k [w_k^{(1)'} X_k' u^{(1)}]^2 \\ &= (1/N^2) \sum_k [b_k^{(1)'} (X_k' X_k)^{-1/2} X_k' u^{(1)}]^2 \end{aligned} \tag{6.11}$$

en posant $b_k^{(1)} = (X_k' X_k)^{1/2} w_k^{(1)}$. La contrainte de norme $\|t_k^{(1)}\| = 1$ se traduit par $\|b_k^{(1)}\| = 1$. Pour $u^{(1)}$ fixée, la valeur optimale de $b_k^{(1)}$ est $b_k^{(1)} = (X_k' X_k)^{-1/2} X_k' u^{(1)} / \|(X_k' X_k)^{-1/2} X_k' u^{(1)}\|$. En reportant cette valeur dans l'expression (6.11) et en remplaçant $u^{(1)}$ par $Yv^{(1)}$, il s'ensuit que :

$$\sum_k cov^2(u^{(1)}, t_k^{(1)}) = (1/N^2) \sum_k v^{(1)'} Y' X_k (X_k' X_k)^{-1} X_k' Y v^{(1)} \tag{6.12}$$

Ce qui permet de conclure que le vecteur $v^{(1)}$ qui maximise ce critère, est le premier vecteur propre normé, associé à la plus grande valeur propre $\lambda^{(1)}$, de la matrice $H = (1/N^2) \sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$. La composante associée à ce vecteur dans l'espace engendré par les variables de Y découle de $u^{(1)} = Yv^{(1)}$. Les composantes partielles sont données par $t_k^{(1)} = X_k w_k^{(1)} = X_k (X_k' X_k)^{-1/2} b_k^{(1)} = P_k u^{(1)} / \|P_k u^{(1)}\|$, où $P_k = X_k (X_k' X_k)^{-1} X_k'$ est le projecteur sur l'espace engendré par les variables de X_k. Les coefficients sont donnés par $a_k^{(1)} = cov(u^{(1)}, t_k^{(1)}) / \sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})} = \|P_k u^{(1)}\| / \sqrt{\sum_l \|P_l u^{(1)}\|^2}$. La composante globale $t^{(1)}$ est donnée par $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)} = \sum_k P_k u^{(1)} / \sqrt{\sum_k \|P_k u^{(1)}\|^2}$.

Il apparaît que la solution du problème considéré est basée sur le premier vecteur propre de $H = (1/N^2) \sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$. Etant donné que les projecteurs P_k sont symétriques et idempotents, il s'ensuit que $H = (1/N^2) \sum_k (P_k Y)' (P_k Y)$. La méthode consiste donc à réaliser une *ACP* du tableau obtenu par concaténation verticale des projections de Y sur les espaces engendrés par les blocs X_k. Cela revient à chercher une direction commune v dans l'espace $\mathfrak{R}^Q$ telle que les nuages associés à $P_k Y$ $(k = 1, \ldots, K)$ aient une inertie moyenne restituée la plus élevée possible. La première valeur propre de la matrice H est égale à $\lambda^{(1)} = (1/N^2) \sum_k \|P_k u^{(1)}\|^2 = (1/N) \sum_k var(P_k u^{(1)})$. Ainsi, cette valeur est d'autant plus grande qu'il existe une direction dans l'espace Y expliquée par les différents tableaux X_k. Naturellement, la contribution relative du tableau X_k dans cette explication peut être évaluée par $\left[(1/N^2)\|P_k u^{(1)}\|^2\right] / \lambda^{(1)}$, et il est aisé de vérifier que cette quantité est égale à $(a_k^{(1)})^2$.

Afin de déterminer les solutions d'ordre deux, la même démarche est effectuée en remplaçant les tableaux X_k et Y par leurs résidus respectifs de la régression sur la première composante globale $t^{(1)}$. Cette procédure peut être répétée plusieurs

fois pour obtenir des composantes globales, ainsi que les composantes partielles associées, d'ordre suivant. Les composantes $(t^{(1)},\ldots,t^{(h)})$ ainsi obtenues peuvent servir à des fins de prédiction, en régressant les variables Y sur celles-ci. Ainsi, la procédure de déflation conduit à l'obtention de composantes globales mutuellement orthogonales, qui, de proche en proche, restituent la variabilité du tableau Y.

Lorsque le tableau X est composé d'un seul bloc, le critère (6.8) montre que l'application de l'*ACPVI* multibloc conduit à l'*ACPVI*. Supposons maintenant que le tableau X soit disposé en autant de blocs qu'il a de variables : $X_1 = [x_1],\ldots,X_P = [x_P]$. Il est clair, à partir du critère (6.9), que l'application de l'*ACPVI* multibloc conduit à la régression *PLS* de Y sur X^*, X^* étant obtenu à partir du tableau X par standardisation des variables. Considérons maintenant le cas de K tableaux $X_1,\ldots,X_K$, et désignons par Y le tableau concaténé $[X_1|\ldots|X_K]$. L'application de l'*ACPVI* multibloc à ce cas de figure vise en définitive à explorer les relations entre les tableaux X_k. A partir du critère (6.10), nous pouvons conclure que dans ce cas, nous sommes conduits à la version modifiée de l'*ACG* présentée paragraphe 6.1.2 page 94.

ACPVI multibloc à résolution itérative

Une autre approche ayant les mêmes objectifs que l'*ACPVI* multibloc est présentée. L'idée est de considérer à nouveau le critère (6.6) page 96, $\sum_q \sum_k cov^2(y_q,t_k)$, qui conduit à des *ACPVI* séparées du tableau Y avec chacun des tableaux X_k. Afin de contraindre les composantes t_k à être liées les unes aux autres, un terme multiplicatif est introduit. De manière précise, il s'agit de maximiser le critère (6.13).

$$\begin{aligned} &var(t^{(1)})\sum_{q=1}^{Q}\sum_{k=1}^{K} cov^2(y_q,t_k^{(1)}) \quad \text{avec} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}, \\ &t_k^{(1)} = X_k w_k^{(1)}, \quad \sum_k a_k^{(1)2} = 1, \quad \|t_k^{(1)}\| = 1 \end{aligned} \tag{6.13}$$

Nous pouvons remarquer que lorsque le tableau X est composé d'un seul bloc, l'application de l'*ACPVI* multibloc à résolution itérative conduit à l'*ACPVI*. De la même façon que pour les méthodes précitées, il est possible d'introduire une composante u, résumé des variables du tableau Y dans le critère à maximiser. En effet, nous pouvons montrer que le critère (6.13) est équivalent au critère (6.14).

$$\begin{aligned} &var(t^{(1)})\sum_{k=1}^{K} cov^2(u^{(1)},t_k^{(1)}) \quad \text{avec} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}, \\ &t_k^{(1)} = X_k w_k^{(1)}, \quad u^{(1)} = Y v^{(1)}, \quad \sum_k a_k^{(1)2} = 1, \quad \|t_k^{(1)}\| = \|v^{(1)}\| = 1 \end{aligned} \tag{6.14}$$

La résolution du critère (6.13) passe par une formulation équivalente de ce critère sous la forme d'un produit de deux formes quadratiques. Au préalable, comme pour la résolution du critère de l'*ACPVI* multibloc, la contrainte de norme $\|t_k\| = 1$ est transformée en $\|b_k\| = 1$ en posant $b_k = (X_k'X_k)^{1/2}w_k$, ce qui donne $t_k = X_k w_k =$

$X_k(X_k'X_k)^{-1/2}b_k = \Pi_k b_k$, en posant $\Pi_k = X_k(X_k'X_k)^{-1/2}$. Le premier terme du critère (6.13) est écrit sous la forme :

$$\begin{aligned}
var(t) &= var(\sum_k a_k t_k) \\
&= \sum_k cov^2(q, t_k) \quad \text{où } q \text{ est la première composante principale normée de } [t_1|\ldots|t_K] \\
&= \sum_k cov^2(q, \Pi_k b_k) \\
&= (1/N^2)\sum_k b_k'\Pi_k' q q'\Pi_k b_k \\
&= (1/N^2)\sum_k b_k' A_k b_k \quad \text{en posant} \quad A_k = \Pi_k' q q'\Pi_k
\end{aligned}$$

Le second terme du critère (6.13) est écrit sous une forme similaire :

$$\begin{aligned}
\sum_k\sum_q cov^2(y_q, t_k) &= \sum_k\sum_q cov^2(y_q, \Pi_k b_k) \\
&= (1/N^2)\sum_k b_k'\Pi_k' Y Y'\Pi_k b_k \\
&= (1/N^2)\sum_k b_k' B_k b_k \quad \text{en posant} \quad B_k = \Pi_k' Y Y'\Pi_k
\end{aligned}$$

Le critère (6.13) à maximiser s'écrit à présent sous la forme (6.15), en définissant $b = [b_1'|\ldots|b_K']'$, ainsi que A et B, les matrices bloc-diagonales ayant comme blocs diagonaux respectifs les blocs $A_k = \Pi_k' q q'\Pi_k$ et $B_k = \Pi_k' Y Y'\Pi_k$.

$$\begin{aligned}
&(b^{(1)'}A^{(1)}b^{(1)})(b^{(1)'}B^{(1)}b^{(1)}) \\
&\text{avec} \quad b^{(1)} = [b_1^{(1)'}|\ldots|b_K^{(1)'}]' \quad \text{et} \quad \|b_k^{(1)}\| = 1
\end{aligned} \tag{6.15}$$

L'écriture du critère à maximiser sous la forme (6.15) montre que celui-ci est un cas particulier du critère proposé par Hanafi et Kiers [2006]. Les auteurs proposent un algorithme itératif général qui permet de le résoudre. La convergence monotone de l'algorithme est garantie par l'application directe du critère de convergence proposé par Hanafi et Kiers [2006]. Une dépendance à l'initialisation est possible. La solution d'ordre un du problème (6.15) est résolue par l'algorithme itératif suivant :

1. Les vecteurs $b^{(1)}$ et $q^{(1)}$ sont initialisés par des vecteurs normés de tailles respectives $(P\times 1)$ et $(N\times 1)$.
2. La dérivée du critère (6.15) par rapport à $b^{(1)}$ est donnée par la matrice $2C^{(1)}b^{(1)}$ où $C^{(1)} = [b^{(1)'}A^{(1)}b^{(1)}]B^{(1)} + [b^{(1)'}B^{(1)}b^{(1)}]A^{(1)}$, tout d'abord calculée à partir des valeurs initiales de $b^{(1)}$ et $q^{(1)}$.
   (a) La matrice $C^{(1)}$ est décomposée en K lignes $C_k^{(1)}$, telles que $C^{(1)} = [C_1^{(1)'}|\ldots|C_K^{(1)'}]'$. A partir des vecteurs $C_k^{(1)}$ sont calculés les vecteurs normés $b_k^{(1)} = C_k^{(1)}b^{(1)}/\|C_k^{(1)}b^{(1)}\|$, avec $b^{(1)'} = [b_1^{(1)'}|\ldots|b_K^{(1)'}]$.

(b) On en déduit les composantes partielles $t_k^{(1)} = X_k(X_k'X_k)^{-1/2}b_k^{(1)}$ pour $k = (1,\ldots,K)$.

(c) La composante $q^{(1)}$ est la première composante principale normée de l'*ACP* du tableau concaténé des composantes partielles $[t_1^{(1)}|\ldots|t_K^{(1)}]$.

(d) La matrice $C^{(1)}$ est recalculée, jusqu'à convergence, à partir des nouvelles valeurs des vecteurs $b^{(1)}$ et $q^{(1)}$.

3. Une fois la convergence obtenue, les composantes globales et partielles, $t^{(1)}$ et $t_k^{(1)}$, sont directement déduites des vecteurs $b^{(1)}$ et $q^{(1)}$: $t_k^{(1)} = X_k(X_k'X_k)^{-1/2}b_k^{(1)}$ et $t^{(1)} = \sum_k a_k^{(1)}t_k^{(1)}$ avec $a_k^{(1)} = cov(q^{(1)}, t_k^{(1)})/\sqrt{\sum_l cov^2(q^{(1)}, t_l^{(1)})}$.

Afin de déterminer les composantes d'ordre deux, la même démarche est effectuée en remplaçant les tableaux X_k et Y par leurs résidus respectifs de la régression sur la première composante globale $t^{(1)}$. Cette procédure est répétée plusieurs fois pour obtenir des composantes globales $(t^{(1)},\ldots,t^{(h)})$, ainsi que les composantes partielles associées. La procédure de déflation conduit à l'obtention de composantes globales mutuellement orthogonales. Les composantes $(t^{(1)},\ldots,t^{(h)})$ obtenues peuvent servir à des fins de prédiction, en régressant les variables Y sur celles-ci.

6.1.4 Méthodes issues de la régression *PLS* pour le cas de $(K+1)$ tableaux

Nous avons souligné que les méthodes présentées dans les paragraphes 6.1.2 et 6.1.3 peuvent être vulnérables à la multicolinéarité à divers degrés : colinéarité parmi les variables Y, au sein des tableaux X_k ou à l'échelle du tableau concaténé X. Pour pallier ce problème, nous considérons des extensions du critère de la régression *PLS* pour le traitement de $(K+1)$ tableaux. Le lecteur intéressé peut se référer à Vivien [2002] pour une revue bibliographique complète sur le sujet. Sont évoquées ici celles basées sur la maximisation d'un critère et dont la résolution est issue d'une décomposition spectrale.

Régression *PLS* multibloc

La méthode *ACPVI* multibloc, présentée dans le paragraphe 6.1.3 page 97, répond aux problématiques de traitement des données d'épidémiologie animale, mais semble sensible au problème de multicolinéarité au sein de chaque tableau X_k du fait des contraintes de norme imposées aux composantes partielles t_k. Nous proposons de considérer le même critère en posant cette fois des contraintes de norme sur les axes associés aux composantes t_k plutôt que sur les composantes t_k elles-mêmes. De manière plus précise, le critère (6.16) à maximiser est :

$$\sum_{q=1}^{Q} cov^2(y_q, t^{(1)}) \quad \text{avec} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}, \quad t_k^{(1)} = X_k w_k^{(1)}, \tag{6.16}$$
$$\sum_{k=1}^{K} a_k^{(1)^2} = 1, \quad \|w_k^{(1)}\| = 1$$

Comme précédemment, nous pouvons montrer que les solutions de ce problème de maximisation (6.16) sont également solutions du problème (6.17).

$$cov^2(u^{(1)}, t^{(1)}) \quad \text{avec} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}, \quad u^{(1)} = Yv^{(1)}, \tag{6.17}$$
$$t_k^{(1)} = X_k w_k^{(1)}, \quad \sum_{k=1}^{K} a_k^{(1)2} = 1, \quad \|w_k^{(1)}\| = \|v^{(1)}\| = 1$$

La solution de ce problème est donnée par $u^{(1)} = Yv^{(1)}$ avec $v^{(1)}$ le premier vecteur propre normé de $(1/N^2)(Y'XX'Y)$ associé à la plus grande valeur propre $\lambda^{(1)}$, et par $t^{(1)} = Xw^{(1)}$ avec $w^{(1)}$ le premier vecteur propre normé de $(1/N^2)(X'YY'X)$ associé à la même valeur propre $\lambda^{(1)}$. Par la suite $w^{(1)}$ est partitionné en blocs normés conformément à la division de X en K blocs : $w^{(1)} = [w_1^{(1)*}|\ldots|w_K^{(1)*}]'$ avec $w_k^{(1)} = w_k^{(1)*}/\|w_k^{(1)*}\|$. Les composantes $t_k^{(1)}$ sont ensuite déterminées par $t_k^{(1)} = X_k w_k^{(1)*}/\|w_k^{(1)*}\|$. Il découle que les coefficients $a_k^{(1)} = \|w_k^{(1)*}\|$ et ainsi que $\sum_k a_k^{(1)2} = \sum_k \|w_k^{(1)*}\|^2 = \|w^{(1)}\|^2 = 1$. Il ressort de ce développement que la solution est basée sur la régression *PLS* des tableaux X concaténé et Y, sans réellement tenir compte de la partition du tableau X en K blocs.

En remplaçant la composante $t^{(1)}$ par son expression $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)}$ dans le critère (6.17), nous avons une écriture intéressante et équivalente du critère à maximiser, sous la forme (6.18).

$$\sum_{k=1}^{K} cov^2(u^{(1)}, t_k^{(1)}) \quad \text{avec} \quad t_k^{(1)} = X_k w_k^{(1)}, \quad u^{(1)} = Yv^{(1)}, \tag{6.18}$$
$$\|w_k^{(1)}\| = \|v^{(1)}\| = 1$$

Cette fois, seules les composantes $t_k^{(1)} = X_k w_k^{(1)}$ et $u^{(1)} = Yv^{(1)}$, associées respectivement aux tableaux X_k et Y sont déterminées, ce qui donne une autre vision de la méthode, qui devient comparable en terme de critère à l'analyse de concordance généralisée et à l'analyse de co-inertie multiple orthogonale détaillées dans le paragraphe suivant. Vivien [2002, p. 143] démontre que la régression *PLS* multibloc (*MPLS*), dans le cas d'un seul tableau Y à expliquer, est aussi basée sur la maximisation des critères (6.17) et (6.18). La régression *PLS* multibloc est initialement présentée par Wold [1984] sous forme d'un algorithme itératif complexe issu de l'algorithme *NIPALS*, détaillé par la suite par Wangen et Kowalski [1988]. Cette méthode permet de relier K tableaux X_k à un (ou plusieurs) tableau(x) Y. Dans le cas d'un seul tableau Y, Westerhuis *et al.* [1998]; Qin *et al.* [2001] puis Vivien [2002, p. 144] montrent que l'algorithme itératif initial est équivalent à l'algorithme qui consiste à effectuer tout d'abord une régression *PLS* de Y sur le tableau concaténé $[X_1|\ldots|X_K]$ pour trouver les composantes globales $t^{(1)} = Xw^{(1)}$ et $u^{(1)} = Yv^{(1)}$. Ce qui revient à dire que $w^{(1)}$ est le premier vecteur propre normé de $(1/N^2)(X'YY'X)$ et que $v^{(1)}$ le premier vecteur propre normé de $(1/N^2)(Y'XX'Y)$, associés à la plus grande valeur propre $\lambda^{(1)}$. Les axes partiels $w_k^{(1)}$ associés à chaque tableau X_k peuvent être calculés soit à partir du découpage du vecteur $w^{(1)}$ en K sous-vecteurs normés, $w_k^{(1)} = w[k]^{(1)}/\|w[k]^{(1)}\|$, soit à partir du vecteur $v^{(1)}$, $w_k^{(1)} = X_k'u^{(1)}/\|X_k'u^{(1)}\|$. Vivien [2002, p. 144] montre que les

coefficients a_k peuvent être calculés par $a_k^{(1)} = cov(u^{(1)}, t_k^{(1)}) / \sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})}$. Tous ces résultats corroborent le développement de la méthode *PLS* multibloc présentée ci-dessus.

Comme pour l'*ACPVI* multibloc, le choix est fait d'effectuer les déflations par rapport aux composantes globales, afin de focaliser davantage sur la restitution de la variabilité du tableau Y [Westerhuis et Coenegracht, 1997]. En effet, ces auteurs montrent que la reconstitution des tableaux X_k en terme d'inertie expliquée est meilleure avec une déflation sur les composantes globales t que sur les composantes partielles t_k.

Analyse de concordance généralisée et analyse de co-inertie multiple orthogonale

La généralisation de l'analyse de concordance (*CONCORg*) au traitement de $(K+1)$ tableaux est proposée par Lafosse et Hanafi [1997]. *CONCORg* est présentée par ses auteurs comme une généralisation dissymétrique de l'analyse inter-batterie (paragraphe 3.1.3 page 53). *CONCORg* est basée sur la maximisation du critère (6.18) de la méthode *PLS* multibloc. Les composantes $(t_1, \ldots, t_K)$ et u sont donc choisies de façon à ce que chacune résume le tableau auquel elle est associée et que la somme des covariances au carré entre chaque composante t_k et la composante u soit la plus grande possible. On peut noter que l'équivalence du critère (6.18) avec le critère (6.17) pour laquelle une composante globale t est déterminée, est démontrée par Lafosse et Hanafi [1997, Prop. 3.1] et Kissita *et al.* [2004, Annexe 6.1]. De ce fait, les solutions d'ordre un de *CONCORg* sont les mêmes que celles de la méthode *MPLS*. Lafosse et Hanafi [1997] indiquent que la composante $t = \sum_k a_k t_k$, moyenne pondérée des composantes t_k, représente un compromis. Les coefficients a_k, proportionnels à $cov(t_k, u)$ pour chaque valeur de k, indiquent les importances respectives des tableaux X_k pour décrire Y [Hanafi et Lafosse, 2001]. Les solutions d'ordre suivant de *CONCORg* diffèrent de celles de la méthode *MPLS* car elles sont basées sur les déflations des tableaux X_k sur leurs axes w_k respectifs. Du fait du choix des déflations, sur les axes plutôt que sur les composantes, *CONCORg* est une méthode plus descriptive et moins prédictive que la méthode *MPLS*. Comme pour l'analyse de concordance (paragraphe 3.1.3 page 53), du fait de l'orthogonalité des axes partiels, il est possible de décomposer l'inertie de chaque tableau X_k en parts concordante, discordante et bruit, vis à vis du tableau Y. Comme l'orthogonalité des axes w_k implique celle des axes v, la décomposition de l'inertie du tableau Y peut être réalisée de façon similaire. La part concordante représente l'information contenue dans X_k qui explique linéairement Y. La part discordante est l'erreur de régression des variables de X_k sur Y. Le bruit est la part d'information pour laquelle il n'y a pas de lien linéaire entre X_k et Y [Hanafi et Lafosse, 2001].

L'analyse de co-inertie multiple orthogonale (*ACIMO*) proposée par Vivien [1999, 2002] est aussi basée sur la maximisation du critère (6.18). La solution d'ordre un est donc la même que celle des méthodes *CONCORg* et *MPLS* [Vivien, 2002]. Les solutions d'ordre suivant de l'*ACIMO* diffèrent car elles sont basées sur les déflations des tableaux X_k sur leurs composantes t_k respectives et du tableau Y sur

la composante u. Les composantes $(t_k^{(1)},\ldots,t_k^{(h)})$, pour $k=(1,\ldots,K)$, et $(u^{(1)},\ldots,u^{(h)})$ sont donc orthogonales par construction, ce qui permet à chaque composante d'expliquer une part différente (des étapes précédentes) de l'inertie du tableau qu'elle résume [Vivien, 2002]. Le choix de réaliser des déflations sur les composantes partielles ne permet pas de représentation simultanée de l'ensemble des variables ou des individus sur des composantes communes, comme c'est le cas pour la méthode *MPLS* avec déflation sur les composantes globales.

6.1.5 Extension de la *latent root regression* au cas de $(K+1)$ tableaux

Dans l'objectif de proposer une méthode adaptée à la description et la prédiction de $(K+1)$ tableaux, robuste à la multicolinéarité des variables explicatives notamment, nous proposons une méthode, qui constitue une extension de la version modifiée de la *latent root regression* présentée au paragraphe 3.1.2 page 51 pour le cas de variables X structurées en K blocs et d'un tableau Y. Comme pour les méthodes présentées précédemment, chaque tableau est résumé par une composante : la composante u résume le tableau Y, la composante globale t le tableau concaténé $X=[X_1|\ldots|X_K]$ et les K composantes partielles t_k, chacun des blocs X_k. Un rôle central est donné aux composantes associées au tableau concaténé X. Cette méthode est appelée *LRR* multibloc [Bougeard *et al.*, 2007]. Les solutions d'ordre un sont obtenues par maximisation de la somme des carrés des covariances entre la composante globale $t^{(1)}$ et chacune des composantes $u^{(1)}$ et $(t_1^{(1)},\ldots,t_K^{(1)})$, ce qui revient à maximiser le critère :

$$cov^2(u^{(1)},t^{(1)})+\sum_{k=1}^{K} cov^2(t_k^{(1)},t^{(1)}) \quad \text{avec} \quad t^{(1)}=Xw^{(1)}, \quad u^{(1)}=Yv^{(1)} \qquad (6.19)$$
$$t_k^{(1)}=X_k w_k^{(1)} \quad \text{et} \quad \|w^{(1)}\|=\|v^{(1)}\|=\|w_k^{(1)}\|=1$$

Le critère peut s'écrire $(1/N^2)\left[(t'^{(1)}Yv^{(1)})^2+\sum_k (t'^{(1)}X_k w_k^{(1)})^2\right]$. Il est facile de montrer que $w^{(1)}$ est le premier vecteur propre normé de la matrice $(1/N^2)(X'YY'X+X'XX'X)$ associé à la plus grande valeur propre $\lambda^{(1)}$. Les composantes $t^{(1)}$, $u^{(1)}$, $t_k^{(1)}$, et les axes partiels $v^{(1)}$ et $w_k^{(1)}$, sont directement calculés à partir de $w^{(1)}$: $t^{(1)}=Xw^{(1)}$, $v^{(1)}=Y't^{(1)}/\|Y't^{(1)}\|$, $u^{(1)}=Yv^{(1)}$, $w_k^{(1)}=X_k't^{(1)}/\|X_k't^{(1)}\|$ et $t_k^{(1)}=X_k w_k^{(1)}$. La solution d'ordre un de la méthode *LRR* multibloc peut être vue comme une analyse de co-inertie multiple [Chessel et Hanafi, 1996] des tableaux $(Y,X_1,\ldots,X_K)$ sous la contrainte que la composante globale soit une combinaison linéaire du tableau concaténé X [Bougeard *et al.*, 2005a].

Les composantes d'ordre deux sont obtenues en considérant les résidus, $X_k^{(1)}$ et $Y^{(1)}$, de la régression des matrices X_k et Y sur la composante globale $t^{(1)}$. La procédure est répétée plusieurs fois pour obtenir les composantes d'ordre suivant. Cette procédure de déflation conduit à l'obtention de composantes orthogonales qui, de proche en proche, restituent la variabilité du tableau Y. Les composantes $(t^{(1)},\ldots,t^{(h)})$ ainsi obtenues peuvent servir à des fins de prédiction, en régressant les variables Y sur celles-ci : $Y=t^{(1)}c^{(1)'}+\cdots+t^{(h)}c^{(h)'}+Y^{(h)}$. Comme les composantes

globales sont des combinaisons linéaires des variables de X : $t^{(1)} = Xw^{*(1)}, \ldots, t^{(h)} = Xw^{*(h)}$, ce qui mène au modèle $Y = X(w^{*(1)}c^{(1)'} + \cdots + w^{*(h)}c^{(h)'}) + Y^{(h)}$.

6.2 Vision synthétique des méthodes liant K tableaux X_k à un tableau Y

6.2.1 Uniformité des critères associées à différentes contraintes

Synthèse des méthodes $(K+1)$-tableaux

Les méthodes permettant d'étudier K tableaux orientés vers un $(K+1)^{ieme}$ sont basées sur un nombre limité de critères, résumés dans le tableau 6.1. Le critère adopté en général consiste à maximiser la somme des carrés des covariances entre les composantes associées aux tableaux X_k et la composante associée à Y : $cov^2(t_k, u)$. Dans ce cas, seules des composantes partielles sont calculées, ce qui ne permet pas d'avoir une vision globale de l'ensemble des variables étudiées. Cependant, l'équivalence avec le problème qui consiste à maximiser $cov^2(t, u)$, avec t combinaison linéaire des composantes partielles t_k, lève cette limite et permet ainsi la comparaison d'un grand nombre de méthodes, présentées initialement par la maximisation de l'un ou l'autre des deux critères. Cette équivalence est démontrée par Lafosse et Hanafi [1997] pour *CONCORg*, par Vivien [2002] pour les méthodes *ACIMO* et *MPLS*, par Kissita *et al.* [2004] pour l'*ACGTR* et dans le paragraphe 6.1.3 page 97 pour l'*ACPVI* multibloc. Une autre vision de la généralisation des critères pour les méthodes $(K+1)$-tableaux est proposée par Hanafi et Lafosse [2001]; Vivien [2002]; Kissita *et al.* [2004], au travers de la maximisation d'un même critère, associée à des métriques différentes (*i.e.* identité ou Mahalanobis).

Choix des déflations

La déflation des tableaux de données sur des axes ou composantes, permet de retirer de ces tableaux l'information extraite à l'étape précédente. Ceci fournit des solutions emboîtées, ce qui est un avantage pour l'interprétation des résultats. De plus, l'inertie des tableaux est ainsi décomposée sur les dimensions étudiées [Vivien, 2002]. Les axes ou composantes sur lesquels est réalisée la déflation sont mutuellement orthogonaux par construction. La première question est d'effectuer la déflation des tableaux soit sur les axes, soit sur les composantes. La déflation sur les axes oriente vers une interprétation plus descriptive et donc moins explicative que celle sur les composantes. Les méthodes ayant recours à ce type de déflation sont plus orientées vers l'étude des individus que vers celle des variables [Vivien, 2002, p. 146]. Notre objectif étant centré sur la description des liens entre variables et sur la prédiction de Y, le choix est fait d'effectuer la déflation des tableaux sur les composantes plutôt que sur les axes.

La seconde question est de réaliser ces déflations sur les composantes globales ou partielles. Vivien [2002, p. 269] choisit d'effectuer les déflations sur les composantes partielles plutôt que globales. L'auteur considère les prédicteurs comme des blocs et postule que l'intérêt des méthodes multiblocs est de proposer un ajustement par

Méthode	Critère à maximiser	Contrainte	Déflation et orthogonalité
LRR multibloc	$cov^2(t,u) + \sum_k cov^2(t,t_k)$	$\|\|w\|\| = \|\|w_k\|\| = \|\|v\|\| = 1$	Déflation de X sur t $(t\perp)$
ACG ss contrainte		$\|\|w\|\| = \|\|t_k\|\| = \|\|u\|\| = 1$	Déflation de X sur t $(t\perp)$
ACGTR		$\|\|t_k\|\| = \|\|u\|\| = 1$	Déflation de X_k sur t_k $(t_k\perp)$ et $(u\perp)$
PLS multibloc	$\sum_k cov^2(u,t_k)$ ou	$\|\|w\|\| = \|\|w_k\|\| = \|\|v\|\| = 1$	Déflation de X sur t $(t\perp)$
CONCORg	$cov^2(t,u)$ avec $t = \sum_k a_k t_k$ et $\sum_k a_k^2 = 1$	$\|\|w\|\| = \|\|w_k\|\| = \|\|v\|\| = 1$	Déflation de X_k sur w_k $(w_k\perp)$ et $(v\perp)$
ACIMO		$\|\|w\|\| = \|\|w_k\|\| = \|\|v\|\| = 1$	Déflation de X_k sur t_k $(t_k\perp)$ et de Y sur u $(u\perp)$ De plus $(w_k\perp)$ et $(v\perp)$
ACPVI multibloc	$\sum_k cov^2(u,t_k)$ ou $cov^2(t,u)$ avec $t = \sum_k a_k t_k$ et $\sum_k a_k^2 = 1$	$\|\|t_k\|\| = \|\|v\|\| = 1$	Déflation de X sur t $(t\perp)$
ACPVI multibloc à résolution iter.	$var(t)\sum_k cov^2(u,t_k)$ avec $t = \sum_k a_k t_k$ et $\sum_k a_k^2 = 1$	$\|\|t_k\|\| = \|\|v\|\| = 1$	Déflation de X sur t $(t\perp)$

TAB. 6.1 – Méthodes permettant de décrire le lien entre K tableaux X_k $(k = 1,\ldots,K)$ et un tableau Y.

bloc de prédicteurs. Westerhuis et Coenegracht [1997]; Westerhuis et Smilde [2001] discutent ce choix dans le cadre de la méthode *PLS* multibloc. Leur stratégie étant basée sur la qualité de la prédiction, ils montrent que le modèle est plus performant quand il est issu de la déflation sur les composantes globales. C'est cette raison qui nous a amené à préconiser d'effectuer les déflations des tableaux sur les composantes globales. En épidémiologie animale, les facteurs de risque des maladies sont souvent répartis dans chacun des blocs de variables. Comme les informations contenues dans chaque bloc sont peu liées ou fonctionnent en covariation, les modélisations partielles ne donnent qu'une image fragmentée et donc peu représentative de la réalité, du lien entre X et Y. L'utilisation de composantes globales fournit des outils pratiques aussi bien pour l'interprétation des résultats (représentations graphiques de toutes les variables) que pour la prédiction.

6.2.2 Apports des méthodes $(K+1)-$tableaux par rapport aux méthodes 2-tableaux

Pré-traitement des données

Le fait d'introduire plusieurs groupes de variables en tant qu'éléments actifs dans une même analyse impose d'équilibrer leur influence *a priori* dans cette analyse [Lebart *et al.*, 2000, p. 344]. Les variables étant centrées et généralement réduites, l'inertie de chaque tableau est égale au nombre de variables du tableau considéré.

Dans le cadre du traitement des données d'épidémiologie animale, la structure des variables explicatives en blocs de taille déséquilibrée n'a pas de sens biologique et peut poser problème pour la sélection des blocs et des variables influençant une maladie (paragraphe 2.2.2 page 39). Cette limite est reprise par la problématique [3] du travail de recherche (paragraphe 2.3 page 41). Une standardisation par facteur d'échelle est proposée de façon à ce que les tableaux $X = [X_1|\ldots|X_K]$ et Y aient la même inertie, égale à un par exemple ; il découle que l'inertie des tableaux X_k pour $k = (1,\ldots,K)$ est égale à $1/K$. Cette standardisation par facteur d'échelle donne ainsi plus de poids au tableau Y dans l'analyse, car ce tableau comprend souvent beaucoup moins de variables que le tableau X. Cette standardisation par facteur d'échelle est préconisée par Wold *et al.* [1996]; Casin [1996]; Westerhuis et Coenegracht [1997]; Vivien [2002]. La standardisation par facteur d'échelle que nous avons adoptée transforme les tableaux, après centrage et réduction des variables, de la façon suivante :

$$X_k \rightarrow X_k / \sqrt{K.inertie(X_k)} \quad \text{puis} \quad X = [X_1|\ldots|X_K]$$
$$Y \rightarrow Y / \sqrt{inertie(Y)}$$

Il faut donc noter qu'à la suite de cette standardisation par facteur d'échelle, les variables X et Y ne sont plus réduites. On note de plus que, le tableau $X = [X_1|\ldots|X_K]$ étant issu de la concaténation des blocs X_k après standardisation par facteur d'échelle, les variances des variables le constituant ne sont pas les mêmes, les variables appartenant à un même bloc ayant une même variance. Cette façon de pré-traiter les données correspond parfaitement aux contraintes des données d'épidémiologie animale : les variables doivent être réduites car elles ont des échelles de mesure différentes. Cependant, un bloc contenant beaucoup de variables a la même influence qu'un groupe en contenant moins.

Influence des blocs dans l'analyse

Dans le cadre de la prédiction d'un tableau Y par K tableaux X_k ($k = 1,\ldots,K$), il est important pour l'utilisateur de mesurer l'influence de chaque bloc X_k dans l'explication de Y. Dans le cas des méthodes *ACGTR*, *ACPVI* multibloc, *ACPVI* multibloc à résolution itérative et *PLS* multibloc, des coefficients a_k normés sont introduits de façon à ce que $t = \sum_k a_k t_k$. Pour les méthodes *ACGTR*, *ACPVI* multibloc et *PLS* multibloc, les coefficients $a_k^{(h)}$ sont calculés pour chaque dimension $h = (1,\ldots,H)$ par la formule (6.20). L'utilisation de la moyenne normée de ces coefficients élevés au carré sur les dimensions $(1,\ldots,h_{opt})$, h_{opt} étant la dimension optimale du modèle liant Y à X, peut être appliquée pour l'évaluation de l'influence des blocs. Cette influence moyenne est nommée $B_k^{(1-h_{opt})}$ (formule 6.21). Il faut noter que, même si ces coefficients sont toujours calculables, ils ne sont pas optimisés par la méthode *LRR* multibloc.

$$a_k^{(h)} = \frac{cov(u^{(h)}, t_k^{(h)})}{\sqrt{\sum_l cov^2(u^{(h)}, t_l^{(h)})}} \tag{6.20}$$

$$B_k^{(1-h_{opt})} = \frac{\sum_{h=1}^{h_{opt}} a_k^{(h)2}}{h_{opt}} \tag{6.21}$$

D'autres critères sont discutés par Vivien [2002, p. 130], mais ils ne nous semblent pas parfaitement adaptés à nos données. L'utilisation de la norme des vecteurs des coefficients $\|\beta_{[k]}\|$ relatifs à chaque tableau X_k rapportée à celle du vecteur $\|\beta\|$ est intéressante. Cependant, elle comporte des limites pour les méthodes sensibles aux quasi-colinéarités car la norme du vecteur des coefficients peut être arbitrairement grande.

Sensibilité à la multicolinéarité

Au niveau conceptuel, l'idée de résumer des tableaux de variables par des composantes devrait suffire à diminuer la sensibilité des résultats à la multicolinéarité des variables au sein de chacun des tableaux. Pourtant, la façon de calculer ces composantes peut être sensible à la multicolinéarité au sein de certains tableaux. Cette sensibilité varie selon les méthodes et peut s'évaluer au travers des matrices à inverser dans les critères à maximiser. A ce titre, les méthodes *LRR* multibloc, *PLS* multibloc, *CONCORg* et *ACIMO* semblent peu sensibles aux multicolinéarités entre variables. Leur résolution ne nécessite aucune inversion de matrice de variance-covariance. A l'inverse, l'*ACG* sous contrainte est à la fois sensible aux multicolinéarités au sein du tableau Y et de chacun des tableaux X_k. l'*ACGTR* est sensible aux multicolinéarités au sein des tableaux X et Y, et l'*ACPVI* multibloc à la multicolinéarité des variables au sein des K tableaux $(X_1, \ldots, X_K)$. L'*ACPVI* multibloc a résolution itérative ayant un mode de résolution complexe, sa sensibilité aux multicolinéarités est évaluée par la suite sur une application. Des recommandations sur l'utilisation de ces méthodes peuvent donc être données par le tableau 6.2.

Multicolinéarité au sein	**Méthodes sensibles**	**Méthodes alternatives**
. . .des tableaux X_k	*ACPVI* mult., *ACG* ss contrainte	*LRR* mult., *PLS* mult. *CONCORg*, *ACIMO*, *ACGTR*
. . .du tableau X	*ACGTR*	*LRR* mult., *PLS* mult. *CONCORg*, *ACIMO*, *ACPVI* mult.
. . .du tableau Y	*ACGTR*, *ACG* ss contrainte	*LRR* mult., *PLS* mult. *CONCORg*, *ACIMO*, *ACPVI* mult.

TAB. 6.2 – Sensibilité des méthodes $(K+1)$-tableaux à la multicolinéarité au sein des différents tableaux de variables.

6.2.3 Choix de la dimension optimale du modèle de régression

Pour définir le nombre optimum de composantes du modèle liant Y à l'ensemble des variables X, la méthode de validation croisée décrite dans le paragraphe 3.2.2 page 55, avec deux échantillons de calibration et de validation, peut être utilisée [Stone, 1974]. Cette méthode permet le calcul de l'erreur moyenne de calibration ($RMSE_C$) et de l'erreur moyenne de validation ($RMSE_V$), calculées à partir des résultats de la validation croisée. Les erreurs $RMSE_C$ et $RMSE_V$, ainsi que les indices $Q^{(h)2}$ et $Q^{(h)2}_{cum}$ qui en sont issus, définis à l'origine dans le cadre de la régression *PLS*, sont adaptables aux modèles multi-tableaux [Vivien, 2002, p. 129]. Il faut noter que pour le cas où les variables Y sont centrées et réduites, le calcul de $RMSE^{(0)}_C = \sum_{q=1}^{Q}\sum_{n=1}^{N} y_{qn}^2 = \sqrt{(N-1)/N}$ [Tenenhaus, 1998, p. 83]. Cependant, dans le contexte considéré ici, le tableau Y est standardisé après le centrage et la réduction des variables (paragraphe 6.2.2), ce qui donne $RMSE^{(0)}_C = \sqrt{(N-1)/(NQ)}$.

Louwerse *et al.* [1999] proposent deux démarches de validation plus spécifiquement adaptées aux méthodes multiblocs (modèle Tucker3 dans cet exemple). Cependant, l'application de ces démarches dans le cadre des méthodes discutées ici suppose que l'on adapte ces méthodes au cas des tableaux comportant des données manquantes.

Chapitre 7

Continuum de méthodes permettant de décrire et relier $(K+1)$ tableaux

7.1 Un continuum pour cadre général aux méthodes liant $(K+1)$ tableaux

7.1.1 Proposition d'un continuum

De la même façon que pour les méthodes liant deux tableaux X et Y (paragraphe 4.1.1 page 59), il est possible de résumer toutes ces méthodes au travers d'un même continuum résumé par le critère (7.1). Le critère à maximiser, ainsi que les contraintes de norme qui lui sont associées, sont modifiés par les variations de trois paramètres. Ce continuum permet de regrouper et d'unifier les méthodes décrites dans le tableau 6.1. Il faut bien sûr noter que cette synthèse n'est valable que pour les solutions d'ordre un, car les déflations sont ensuite spécifiques à chaque méthode.

$$
\begin{aligned}
&\alpha cov^2(u^{(1)}, t^{(1)}) + (1-\alpha)\sum_k cov^2(t_k^{(1)}, t^{(1)}) \\
&\text{avec} \quad t^{(1)} = Xw^{(1)}, \quad u^{(1)} = Yv^{(1)}, \quad t_k^{(1)} = X_k w_k^{(1)} \\
&\gamma_1 \|w_k^{(1)}\|^2 + (1-\gamma_1)\|t_k^{(1)}\|^2 = 1, \quad \gamma_2 \|v^{(1)}\|^2 + (1-\gamma_2)\|u^{(1)}\|^2 = 1 \\
&0 \leq \alpha \leq 1, \quad 0 \leq \gamma_1 \leq 1, \quad 0 \leq \gamma_2 \leq 1
\end{aligned}
\tag{7.1}
$$

Il est possible de présenter la plupart des méthodes traitant $(K+1)$ tableaux dans un même espace à trois dimensions (figure 7.1) en se basant sur les variations des trois paramètres α, γ_1 et γ_2, dont le rôle est détaillé dans le paragraphe 4.1.2 page 61. Ce résumé n'est valide que pour les solutions d'ordre un. Le positionnement des méthodes $(K+1)-$tableaux de la figure 7.1 peut être directement mis en regard avec celui des méthodes liant 2 tableaux, présenté dans le paragraphe 4.1.1 page 59.

Comme pour le cas des méthodes liant deux tableaux X et Y (paragraphe 4.1.4 page 65), le paramètre α ajuste la méthode à l'objectif du traitement statistique, plus ou moins orienté vers l'explication du tableau Y. Les paramètres γ_1 et γ_2 modifient

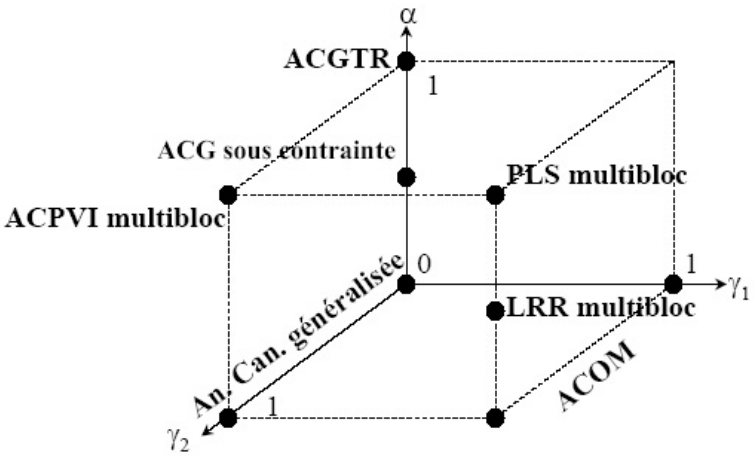


FIG. 7.1 – Illustration des cas particuliers du continuum généralisant les principales méthodes liant K tableaux X_k pour $k = (1,\ldots,K)$ à un tableau Y (solutions d'ordre un).

les contraintes de norme associées à chaque tableau et ajustent ainsi la méthode pour pallier le problème de multicolinéarités des variables au sein de chaque bloc.

Selon les valeurs des trois paramètres du continuum, différents cas particuliers apparaissent. Pour le cas où le paramètre α est nul (pas d'orientation vers l'explication d'un tableau Y), on retrouve soit l'analyse canonique généralisée (normalisation des composantes), soit l'analyse de co-inertie multiple, ou *ACOM* [Chessel et Hanafi, 1996] (normalisation des axes). Les deux méthodes intermédiaires proposées, *latent root regression* multibloc et analyse canonique généralisée sous contrainte, se placent dans le plan intermédiaire ($\alpha = 1/2$). Du fait du critère maximisé, ces deux méthodes apparaissent à mi-chemin entre les analyses K tableaux ($\alpha = 0$) et les analyses $(K+1)$ tableaux ($\alpha = 1$). Pour le cas ($\alpha = 1$), les méthodes *ACGTR*, *PLS* multibloc et *ACPVI* multibloc, sont respectivement basées sur des extensions de l'*ACG*, de la régression *PLS* et de l'*ACPVI* pour le cas de K tableaux X_k et d'un tableau à expliquer Y.

7.1.2 Sélection des continuums à explorer dans le cadre du traitement des données d'épidémiologie animale

Comme pour le cas des méthodes liant deux tableaux X et Y (paragraphe 4.1.4 page 65), l'idée d'un continuum unifiant un grand nombre de méthodes traitant $(K+1)$ tableaux est intéressante du point de vue théorique, chaque paramètre ayant un rôle clair qui facilite son interprétation. Cependant, du point de vue pratique, l'exploration d'un continuum à trois dimensions (voire quatre en incluant le nombre de dimension h du modèle) semble peu réaliste. Quelques continuums, basés sur la variation d'un seul paramètre, répondant aux problématiques relatives aux données d'épidémiologie animale exposées dans le paragraphe 2.3 page 41, sont donc sélectionnés.

Les deux continuums explorés pour le cas de deux tableaux (paragraphe 4.2 page 66) sont étendus au cas de $(K+1)$ tableaux. Le premier continuum exploré

est donc une extension du continuum *LRR* (paragraphe 4.2.1 page 66), basé sur les variations du paramètre α. Afin de mieux comprendre l'influence des contraintes de norme dans le cadre des méthodes traitant $(K+1)$ tableaux, un continuum parallèle basé lui aussi sur les variations du paramètre α est exploré autour de l'analyse canonique généralisée sous contrainte. Le dernier continuum exploré dans le cadre du traitement de deux tableaux (continuum $ACPVI-PLS$, paragraphe 4.2.2 page 68) est lui aussi étendu au cas multibloc.

7.2 Continuums explorés dans le cadre de $(K+1)$ tableaux

7.2.1 Continuum *LRR* multibloc

Le continuum *LRR* multibloc est une extension du continuum *LRR* adapté au traitement de deux tableaux (paragraphe 4.2.1 page 66). Il est basé sur les variations du paramètre α au travers de la maximisation du critère (7.2) [Bougeard *et al.*, 2005b].

$$\begin{aligned}
&(1-\alpha)\sum_{k=1}^{K} cov^2(t_{k,\alpha}^{(1)}, t_{\alpha}^{(1)}) + \alpha cov^2(u_{\alpha}^{(1)}, t_{\alpha}^{(1)}) \\
&\text{avec} \quad t_{\alpha}^{(1)} = Xw_{\alpha}^{(1)}, \quad t_{k,\alpha}^{(1)} = X_k w_{k,\alpha}^{(1)}, \quad u_{\alpha}^{(1)} = Yv_{\alpha}^{(1)} \\
&\|w_{\alpha}^{(1)}\| = \|w_{k,\alpha}^{(1)}\| = \|v_{\alpha}^{(1)}\| = 1 \quad \text{et} \quad 0 \leq \alpha \leq 1
\end{aligned} \tag{7.2}$$

La solution est donnée par $w_{\alpha}^{(1)}$ premier vecteur propre de la matrice $(1/N^2)[\alpha X'YY'X + (1-\alpha)X'XX'X]$ associé à la plus grande valeur propre $\lambda_{\alpha}^{(1)}$. Les axes partiels sont calculés à partir de la composante $t_{\alpha}^{(1)} = Xw_{\alpha}^{(1)} : v_{\alpha}^{(1)} = Y't_{\alpha}^{(1)}/\|Y't_{\alpha}^{(1)}\|$ et $w_{k,\alpha}^{(1)} = X_k't_{\alpha}^{(1)}/\|X_k't_{\alpha}^{(1)}\|$. Les solutions d'ordre suivant sont obtenues par remplacement des tableaux X et Y par leurs résidus de la régression de X et Y sur la première composante $t_{\alpha}^{(1)}$ dans le critère à maximiser. La figure 7.2 illustre la famille de méthodes explorées par le continuum *LRR* multibloc. En faisant varier le paramètre α, le continuum explore les solutions comprises entre l'analyse de co-inertie multiple [Chessel et Hanafi, 1996] des tableaux $(X_1, \ldots, X_K)$ pour $(\alpha = 0)$, la méthode *LRR* multibloc appliquée à $(K+1)$ tableaux $(Y, X_1, \ldots, X_K)$ pour $(\alpha = 1/2)$ et la régression *PLS* multibloc pour $(\alpha = 1)$.

7.2.2 Continuum *ACG* sous contrainte

Le continuum associé à l'analyse canonique généralisée sous contrainte (paragraphe 6.1.2 page 94) peut être mis en regard du continuum *LRR* multibloc. Le critère à maximiser est identique, seules les contraintes de normes diffèrent. Pour le continuum *ACG* sous contrainte, les contraintes sont posées sur les composantes t_k et u et non pas sur les axes w_k et v comme pour le continuum *LRR* multibloc. Pour ces continuums par contre, une contrainte de norme sur l'axe w est posée, afin de limiter la sensibilité aux quasi-colinéarités entre les variables du tableau concaténé X. La comparaison des résultats de ces deux continuums permet d'apprécier l'influence des contraintes de normes sur les résultats donnés par les méthodes. Le

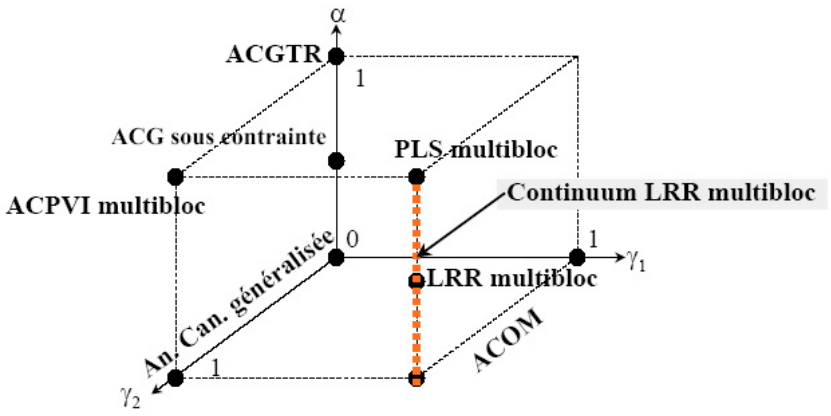


FIG. 7.2 – Illustration du domaine exploré par le continuum *LRR* multibloc.

critère du continuum *ACG* sous contrainte est basé sur les variations du paramètre α au travers la maximisation du critère (7.3).

$$
\begin{aligned}
&(1-\alpha)\sum_{k=1}^{K} cov^2(t_{k,\alpha}^{(1)}, t_\alpha^{(1)}) + \alpha cov^2(u_\alpha^{(1)}, t_\alpha^{(1)}) \\
&\text{avec} \quad t_\alpha^{(1)} = Xw_\alpha^{(1)}, \quad t_{k,\alpha}^{(1)} = X_k w_{k,\alpha}^{(1)}, \quad u_\alpha^{(1)} = Yv_\alpha^{(1)} \\
&\|w_\alpha^{(1)}\| = \|t_{k,\alpha}^{(1)}\| = \|u_\alpha^{(1)}\| = 1 \quad \text{et} \quad 0 \le \alpha \le 1
\end{aligned}
\tag{7.3}
$$

La solution de ce problème est donnée par $w_\alpha^{(1)}$ premier vecteur propre de la matrice $(1/N^2)\left[\alpha X'Y(Y'Y)^{-1}Y'X + (1-\alpha)X'\sum_k (X_k(X_k'X_k)^{-1}X_k')X\right]$ associé à la plus grande valeur propre $\lambda_\alpha^{(1)}$. Les composantes partielles sont calculées à partir de la projection de la composante $t_\alpha^{(1)} = Xw_\alpha^{(1)}$ sur les espaces associés aux autres tableaux. Ceci conduit à $u_\alpha^{(1)} = [Y(Y'Y)^{-1}Y't_\alpha^{(1)}]/\|Y(Y'Y)^{-1}Y't_\alpha^{(1)}\|$ et $t_{k,\alpha}^{(1)} = [X_k(X_k'X_k)^{-1}X_k't_\alpha^{(1)}]/\|X_k(X_k'X_k)^{-1}X_k't_\alpha^{(1)}\|$. Les solutions d'ordre suivant sont obtenues par remplacement des tableaux X et Y par leurs résidus de la régression sur la première composante $t_\alpha^{(1)}$ dans le critère à maximiser. La figure 7.3 illustre la famille de méthodes explorées par le continuum *ACG* sous contrainte. En faisant varier le paramètre α, le continuum explore les solutions comprises entre l'analyse canonique généralisée des tableaux $(X_1, \ldots, X_K)$ pour $(\alpha = 0)$, l'analyse canonique généralisée sous contrainte appliquée aux $(K+1)$ tableaux $(Y, X_1, \ldots, X_K)$ pour $(\alpha = 1/2)$ et l'*ACGTR* pour $(\alpha = 1)$. Les solutions d'ordre supérieur à un diffèrent des résultats issus des méthodes *ACG* et *ACGTR* car les déflations de ces méthodes sont réalisées respectivement sur les composantes partielles t_k, ainsi que u et t_k.

7.2.3 Continuum $ACPVI - PLS$ multibloc

La méthode *ACPVI* multibloc est basée, pour la détermination du vecteur v associé au tableau Y, sur la décomposition spectrale de la matrice $\sum_k Y'X_k(X_k'X_k)^{-1}X_k'Y$ et la régression *PLS* multibloc sur la décomposition spectrale de la matrice $(Y'XX'Y) =$

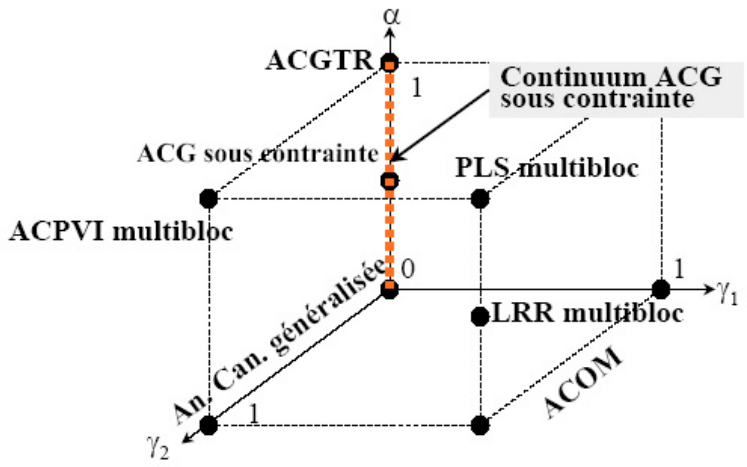


FIG. 7.3 – Illustration du domaine exploré par le continuum *ACG* sous contrainte.

$\sum_k Y'(X_k'X_k)Y$. Il apparaît donc que la méthode *PLS* multibloc est basée sur une contraction des matrices $(X_k'X_k)^{-1}$ vers les matrices identités I_{p_k} pour $k = (1,\ldots,K)$. Le continuum détaillé ici établit un lien entre l'*ACPVI* multibloc et la méthode *PLS* multibloc. Le continuum *ACPVI* – *PLS* multibloc est une extension du continuum *ACPVI* – *PLS*, adaptée au traitement de deux tableaux (paragraphe 4.2.2 page 68). Il est basé sur les variations du paramètre γ_1 à travers la maximisation du critère (7.4).

$$cov^2(u_{\gamma_1}^{(1)}, t_{\gamma_1}^{(1)}) \quad \text{avec} \quad t_{\gamma_1}^{(1)} = \sum_{k=1}^{K} a_{k,\gamma_1}^{(1)} t_{k,\gamma_1}^{(1)}, \quad t_{k,\gamma_1}^{(1)} = X_k w_{k,\gamma_1}^{(1)}, \quad u_{\gamma_1}^{(1)} = Y v_{\gamma_1}^{(1)} \tag{7.4}$$
$$\sum_{k=1}^{K} a_{k,\gamma_1}^{(1)2} = 1, \quad \|v_{\gamma_1}^{(1)}\| = 1, \quad \gamma_1 \|w_{k,\gamma_1}^{(1)}\|^2 + (1-\gamma_1)\|t_{k,\gamma_1}^{(1)}\|^2 = 1, \quad 0 \le \gamma_1 \le 1$$

Pour résoudre la maximisation de ce critère, un changement de variable $b_{k,\gamma_1}^{(1)} = [\gamma_1 I + (1-\gamma_1)(X_k'X_k)]^{1/2} w_{k,\gamma_1}^{(1)}$ est opéré, qui permet d'obtenir $\|b_{k,\gamma_1}^{(1)}\| = 1$. Il s'ensuit que $t_{k,\gamma_1}^{(1)} = X_k[\gamma_1 I + (1-\gamma_1)(X_k'X_k)]^{-1/2} b_{k,\gamma_1}^{(1)} = X_k^* b_{k,\gamma_1}^{(1)}$ avec $X_k^* = X_k[\gamma_1 I + (1-\gamma_1)(X_k'X_k)]^{-1/2}$. La solution de ce problème est donnée en considérant $v_{\gamma_1}^{(1)}$ comme le vecteur propre de la matrice $(1/N^2)\sum_k Y'X_k^* X_k^{*'} Y = (1/N^2)\sum_k Y'X_k[(1-\gamma_1)X_k'X_k + \gamma_1 I]^{-1} X_k'Y$, associé à la plus grande valeur propre $\lambda_{\gamma_1}^{(1)}$. Les axes associés aux tableaux X_k sont calculés à partir de la composante $u_{\gamma_1}^{(1)} = Y v_{\gamma_1}^{(1)}$: $w_{k,\gamma_1}^{(1)} = [(1-\gamma_1)(X_k'X_k) + \gamma_1 I]^{-1} X_k' u_{\gamma_1}^{(1)} / \|[(1-\gamma_1)(X_k'X_k) + \gamma_1 I]^{-1/2} X_k' u_{\gamma_1}^{(1)}\|$. De ces axes, sont déduits les composantes $t_{k,\gamma_1}^{(1)} = X_k w_{k,\gamma_1}^{(1)}$ ainsi que les coefficients $a_{k,\gamma_1}^{(1)} = cov(u_{\gamma_1}^{(1)}, t_{k,\gamma_1}^{(1)}) / \sqrt{\sum_l cov^2(u_{\gamma_1}^{(1)}, t_{l,\gamma_1}^{(1)})}$. La composante globale est ensuite calculée grâce à $t_{\gamma_1}^{(1)} = \sum_{k=1}^{K} a_{k,\gamma_1}^{(1)} t_{k,\gamma_1}^{(1)}$. Les solutions d'ordre suivant sont obtenues par remplacement des tableaux X et Y par leurs résidus de la régression sur la première composante $t_\gamma^{(1)}$ dans le critère à maximiser. La figure 7.4 illustre la famille de méthodes explorées par le continuum. En faisant varier le paramètre γ_1, le continuum explore les solutions comprises entre l'*ACPVI* multibloc des tableaux

$(Y, X_1, \ldots, X_K)$ pour $(\gamma_1 = 0)$ et la méthode *PLS* multibloc pour $(\gamma_1 = 1)$.

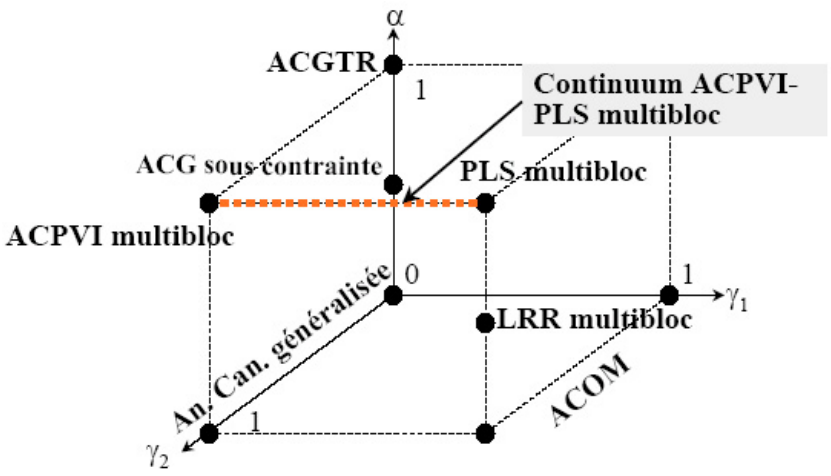


FIG. 7.4 – Illustration du domaine exploré par le continuum *ACPVI – PLS* multibloc.

7.2.4 Sélection des paramètres optimaux des continuums

Comme pour le cas des continuums s'appliquant au traitement de deux tableaux, deux paramètres sont à estimer : le paramètre du continuum (α ou γ_1 selon le continuum étudié) et le nombre h de composantes $(t^{(1)}, \ldots, t^{(h)})$ à retenir dans le modèle expliquant Y par X. La même méthode de validation croisée, détaillée dans le paragraphe 4.2.3 page 73, est donc utilisée.

Chapitre 8

Application au traitement de données d'épidémiologie animale organisées en $(K+1)$ tableaux

8.1 Données et problématique

Les données sont issues d'une enquête analytique portant sur la maladie de l'amaigrissement du porcelet [Rose *et al.*, 2003a]. L'un des principaux facteurs infectieux de cette maladie est le circovirus *PCV*2. Le tableau de données comporte 158 élevages sur lesquels sont mesurées 36 variables organisées en cinq blocs décrits dans le tableau 8.1. Le tableau Y, composé de trois variables, mesure la proportion d'animaux ayant réagi à l'infection par le virus *PCV*2 (truies, porcs charcutiers et porcelets). Les variables X sont organisées en quatre tableaux : X_1 relatif aux mesures de bio-sécurité et d'hygiène, X_2 reflétant la conduite d'élevage, X_3 lié à la structure de l'élevage et X_4 relatif aux co-facteurs infectieux et vaccins. Les variables qualitatives ont été codées selon un codage disjonctif complet. Le premier objectif de l'étude est de décrire les liens entre les variables et entre les blocs de variables, et de déterminer les variables permettant de différencier les élevages. Le second objectif est à la fois de déterminer, parmi les variables de X, celles qui sont facteurs de risque de la prévalence des séropositivités des élevages au circovirus *PCV*2 et de mesurer l'influence des blocs de variables X_k ($k = 1, \ldots, 4$) dans l'explication de cette séropositivité des élevages au circovirus *PCV*2. Les variables, ayant des unités de mesure différentes, sont centrées et réduites. Le nombre de variables différant nettement d'un bloc à un autre, ceux-ci sont standardisés par un facteur d'échelle suivant la procédure décrite dans le paragraphe 6.2.2 page 106.

Les méthodes développées pouvant être sensibles à la multicolinéarité au sein de chacun des tableaux, un indice de conditionnement maximal (décrit dans le paragraphe 4.2.2 page 70) relatif à chaque tableau est calculé et reporté dans le tableau 8.2. Il ne semble pas y avoir de multicolinéarité majeure dans ce jeu de données car l'indice de conditionnement maximal de chaque tableau est inférieur au seuil de 20 [Erkel-Rousse, 1995].

Bloc	Variable	Description
Y	CIRCOPS	Proportion de porcelets ayant réagi à l'infection par le *PCV*2 en post-sevrage
	CIRCOPC	Proportion de porcs ayant réagi à l'infection par le *PCV*2 en engraissement
	CIRCOTR	Proportion de truies ayant réagi à l'infection par le *PCV*2
X_1	MAVPOR	Sens de circulation des animaux (marche en avant)
	MAVHOM	Sens de circulation des hommes (marche en avant)
	PEDILUV	Utilisation d'un pédiluve dans chaque salle de l'élevage
	AIGJET	Utilisation d'une aiguille jetable par truie pour les vaccins
	DETERENG	Détersion supplémentaire de la salle d'engraissement après lavage
	VIDFOSEnon	Vidange de la fosse (partielle *vs* pas de vidange)
	VIDFOSEtot	Vidange de la fosse (partielle *vs* totale)
	DESINFGES	Désinfection des stalles de gestation
	LAVTRUIM	Lavage des truies à l'entrée en maternité
	DVSM	Durée du vide sanitaire en maternité
X_2	GESBANDEmel	Séparation physique des bandes de truies gestantes (externe *vs* mélange)
	GESBANDEsep	Séparation physique des bandes de truies gestantes (externe *vs* séparées)
	GESCOCHTmel	Position des cochettes au sein des travées (externe *vs* mélange)
	GESCOCHTsep	Position des cochettes au sein des travées (externe *vs* séparées)
	NOUPRES	Présence d'une nursery pour une partie de la bande
	AGECAST	Age des porcelets mâles à la castration
	TXREN	Taux de renouvellement du troupeau de truies
	AGESEV	Age des porcelets au sevrage
	PSNPOR	Nombre de portées par case en post-sevrage
	QUARBAND	Nombre de lots par salle en quarantaine
	DURQUAR	Durée moyenne de la quarantaine (en semaines)
	NBVER	Nombre moyen de verrats introduits par an
X_3	NBENG	Nombre d'élevages engraisseurs dans un rayon de 2 km autour de l'élevage
	NBNAIENG	Nombre d'élevages naisseurs-engraisseurs dans un rayon de 2 km autour de l'élevage
	CLOISON	Cloisons entre les préfosses en engraissement
	ALIMENGnourri	Type d'alimentation en engraissement (soupe *vs* nourrisoupe)
	ALIMENGsec	Type d'alimentation en engraissement (soupe *vs* sec)
	SURFENG	Surface des cases en engraissement
	LOCQUAR	Type de locaux en quarantaine (semi-claustration ou claustration)
	FOSMAT	Profondeur des préfosses en maternité
X_4	SDRP	Vaccination des truies contre le virus *SDRP*
	PARVOQG	Utilisation du même antigène contre le parvovirus en quarantaine et lors de la gestation
	PARVOCO	Proportion de cochettes ayant réagi positivement à l'infection au parvovirus

TAB. 8.1 – Description des variables et des blocs de variables.

8.2 Description de tableaux structurés en blocs

8.2.1 Interprétation des composantes

Evolution des inerties expliquées par les composantes globales

La figure 8.1 illustre le pourcentage cumulé des inerties expliquées par les composantes globales $(t^{(1)}, \ldots, t^{(h)})$. La restitution de l'inertie par les composantes globales n'est pas la même pour chacune des méthodes. Par exemple, pour restituer 50% de l'inertie, la méthode *LRR* multibloc nécessite 9 composantes, la *PLS* multibloc 12

Tableau	Y	X	X_1	X_2	X_3	X_4
Indice η	2.02	8.49	2.88	3.74	4.13	1.30

TAB. 8.2 – Indice de conditionnement maximal relatif à chaque tableau pour le jeu de données sur le cirocvirus *PCV*2.

composantes, l'*ACPVI* multibloc 22 composantes et l'*ACPVI* multibloc itérative 23 composantes. Si l'objectif est de décrire les données, une méthode résumant l'information sur un nombre réduit de composantes est préférée (*LRR* multibloc sur cet exemple).

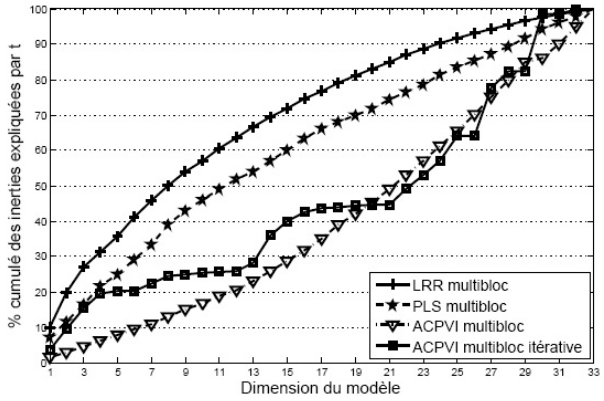


FIG. 8.1 – Pourcentage cumulé des inerties expliquées par les composantes globales $(t^{(1)}, \ldots, t^{(h)})$. Comparaison des résultats des méthodes *LRR* multibloc, *PLS* multibloc, *ACPVI* multibloc et *ACPVI* multibloc itérative.

La variance de la première composante globale $t^{(1)}$ peut être considérée comme une mesure de stabilité de cette composante, étant entendu que lorsque cette variance est relativement faible, nous pourrions être en mesure de supposer qu'elle reflète du bruit. La figure 8.2 illustre ces résultats. L'analyse de co-inertie multiple, ou *ACOM* (paragraphe 7.1.1 page 111), n'est pas orientée vers l'explication du tableau Y ; elle apparaît plus stable que les autres méthodes. Sur cet exemple, la méthode *LRR* multibloc dont le critère est intermédiaire entre l'explication de X et de Y donne des résultats proches de ceux de l'*ACOM*. Les extensions multiblocs de l'*ACPVI* et de la régression *PLS* comportent la même évolution de la variance de $t^{(1)}$ que ceux de l'*ACPVI* et la régression *PLS* pour le cas de deux tableaux : la méthode *PLS* multibloc apparaît plus stable que l'*ACPVI* multibloc.

Evolution de l'inertie des tableaux expliquée par les composantes globales

La figure 8.3 représente les parts d'inertie des tableaux X et Y expliquées par les composantes globales. le tableau X est expliqué de façon à peu près équivalente par les quatre méthodes, même si la méthode *LRR* multibloc en donne une meilleure explication. Pour ce qui est de l'explication du tableau Y par les composantes globales, les résultats des quatre méthodes diffèrent nettement. L'*ACPVI* multibloc puis la régression *PLS* multibloc donnent une bonne explication du tableau Y sur

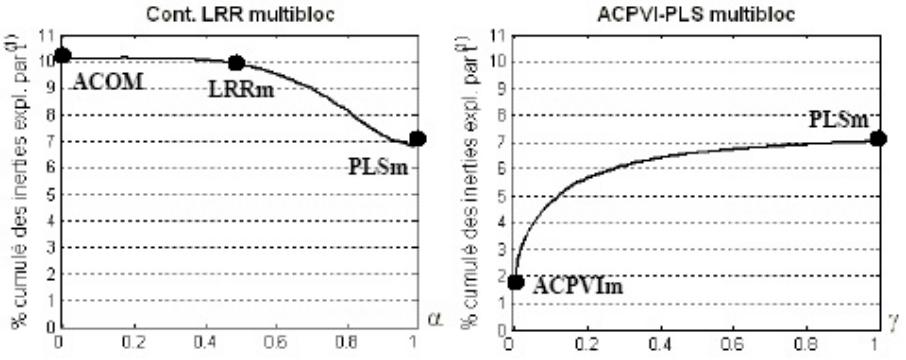


FIG. 8.2 – Comparaison des inerties expliquées par la composante $t^{(1)}$ pour les continuums *LRR* mutibloc et $ACPVI - PLS$ multibloc. Les cas particuliers de ces continuums sont aussi indiqués.

les premières composantes. Ces deux méthodes ont un effet un critère orienté vers l'optimisation du lien entre X et Y. Les méthodes *LRR* multibloc et *ACPVI* à résolution itérative expliquent moins bien le tableau Y. La méthode *PLS* multibloc fournit, sur cet exemple, un compromis intéressant entre une bonne explication du tableau Y et une explication correcte du tableau X.

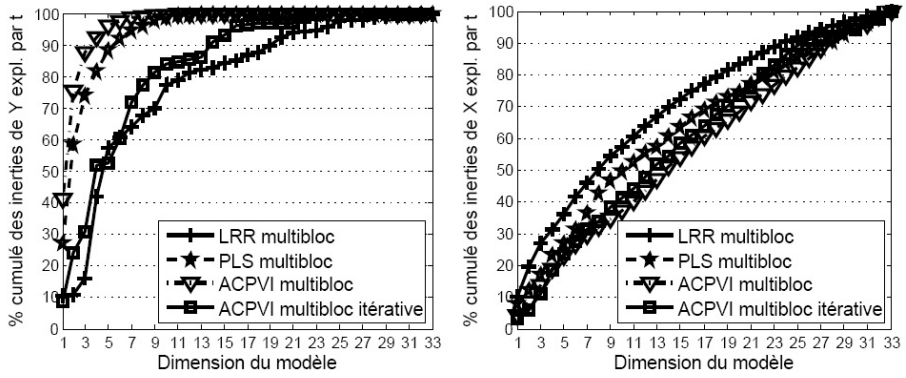


FIG. 8.3 – Pourcentage cumulé des inerties des tableaux X et Y expliquées par les composantes globales. Comparaison des résultats des méthodes *LRR* multibloc, *PLS* multibloc, *ACPVI* multibloc et *ACPVI* multibloc itérative.

L'inertie du tableau Y expliquée par la première composante globale $t^{(1)}$ peut être considérée comme une mesure de la qualité d'ajustement du modèle aux données (figure 8.4). La régression *PLS* multibloc, dont le critère est plus orienté vers l'explication du tableau Y, apparaît sur cet exemple comme ayant une meilleure qualité qualité d'ajustement que l'*ACOM*. Les contraintes de norme ont aussi une influence sur la qualité qualité d'ajustement de la méthode. La comparaison de la qualité qualité d'ajustement de l'*ACPVI* multibloc à celle de la régression *PLS* multibloc montre que le fait de poser des contraintes sur les composantes associées aux tableaux X_k (*ACPVI* multibloc) plutôt que sur les axes (régression *PLS* multibloc) augmente, sur cet exemple, l'orientation de la méthode vers l'explication de Y.

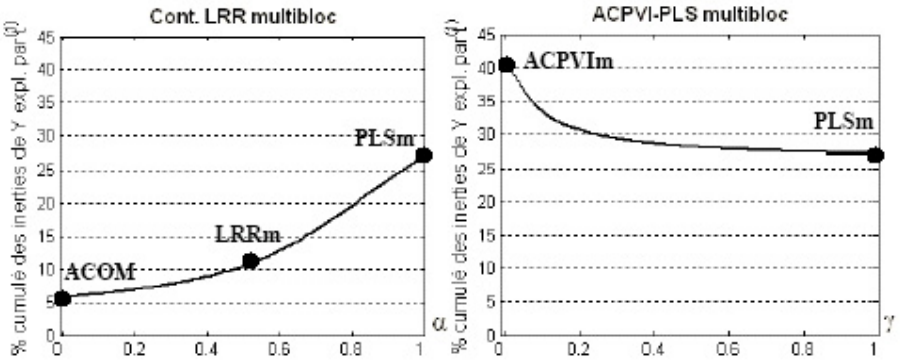


FIG. 8.4 – Comparaison des inerties du tableau Y expliquées par la composante $t^{(1)}$ pour les continuums *LRR* mutibloc et $ACPVI - PLS$ multibloc. Les cas particuliers de ces continuums sont aussi indiqués.

L'explication du tableau X par les composantes globales est comparable d'une méthode à une autre (figure 8.3), mais l'importance des tableaux X_k ($k = 1, \ldots, 4$) dans la construction de ces composantes n'est pas la même. La figure 8.5 illustre ces différences. Les tableaux X_1, X_2 et dans une moindre mesure X_3 sont expliqués de façon comparable par les quatre méthodes, même si les composantes de la méthode *LRR* multibloc expliquent un peu mieux les tableaux X_1, X_2 et surtout X_3. De nettes différences apparaissent pour l'explication du tableau X_4 par les composantes globales. Les méthodes *LRR* multibloc et *PLS* multibloc tiennent nettement plus compte du tableau X_4 pour la construction des composantes globales que les autres méthodes, notamment pour les premières dimensions.

8.2.2 Représentation factorielle

Interprétation de la carte des variables

Les variables de X et de Y peuvent être représentées sur un système formé par les composantes globales t, orthogonales mutuellement par construction. Les figures 8.6 et 8.7 illustrent ces représentations sur le plan des deux premières dimensions ($t^{(1)}$, $t^{(2)}$) pour les méthodes *LRR* multibloc, *PLS* multibloc, *ACPVI* multibloc et *ACPVI* multibloc itérative. Sur les cartes factorielles des méthodes *ACPVI* multibloc et *PLS* multibloc, les variables Y relatives à la proportion de truies (*CIRCOTR*) et de porcelets (*CIRCOPS*) séropositifs au virus *PCV*2 sont liées et non corrélées à la proportion de porcs à l'engrais séropositifs (*CIRCOPC*). Cette image semble proche de la réalité des faits en élevage. En effet, les truies et les porcelets sont élevés ensemble en maternité (au préalable de la prise de sang qui a lieu lorsque les porcelets sont en post-sevrage, séparés de leur mère), ce qui explique que leurs proportions de séroprévalence soient comparables. Les porcs à l'engrais, plus âgés, ont perdu trace de l'immunité colostrale reçue de la truie. Ils sont élevés dans d'autres bâtiments et sont donc soumis à un milieu différent. La méthode *LRR* multibloc fournit des résultats la différenciant nettement des autres méthodes : les trois variables Y sont bien expliquées sur ce plan et essentiellement par les variables

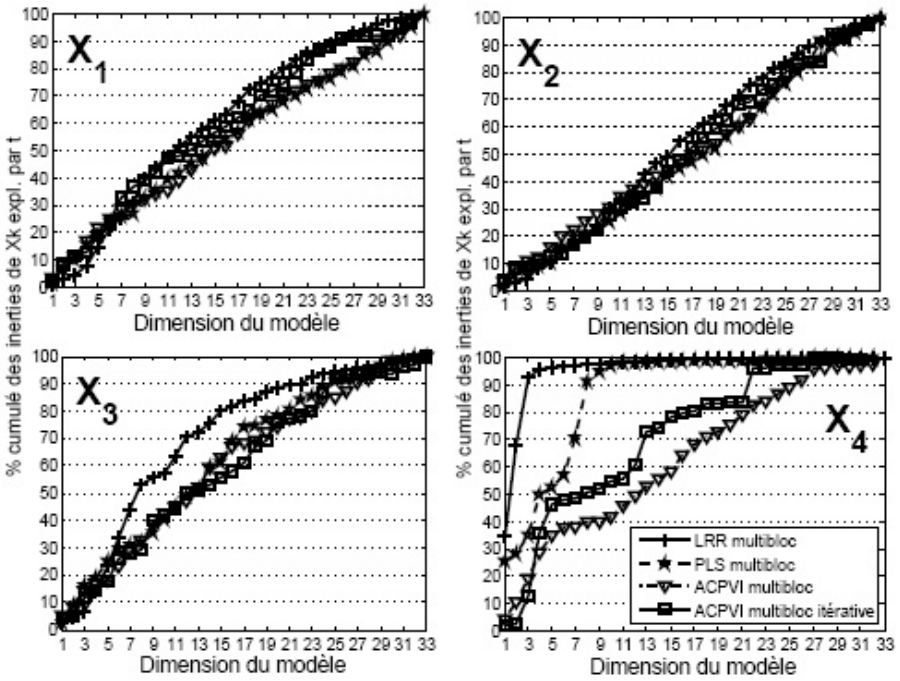
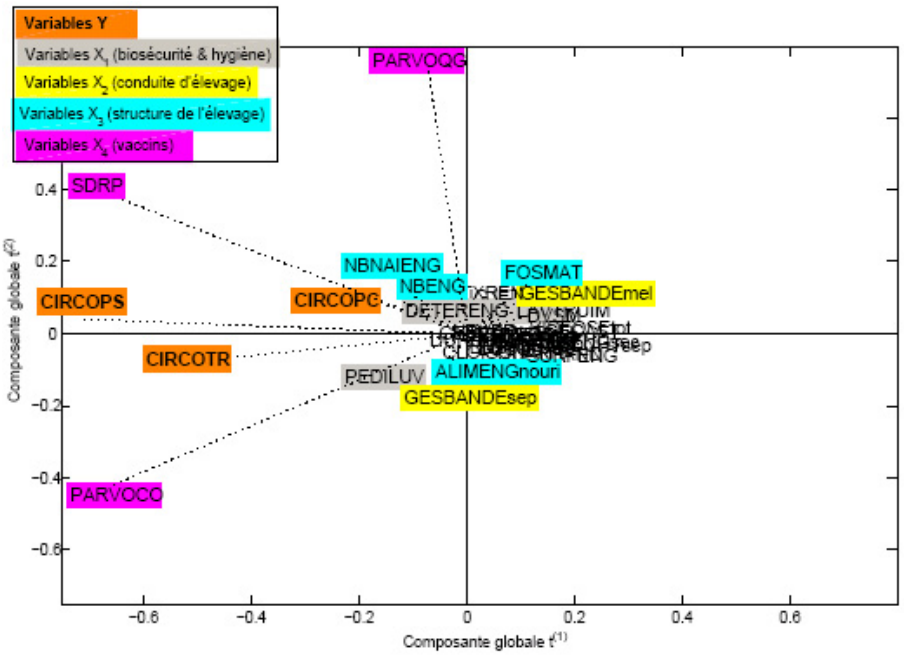


FIG. 8.5 – Pourcentage cumulé des inerties des tableaux X_k ($k = 1, \ldots, 4$) expliquées par les composantes globales. Comparaison des résultats des méthodes *LRR* multibloc, *PLS* multibloc, *ACPVI* multibloc et *ACPVI* multibloc itérative.

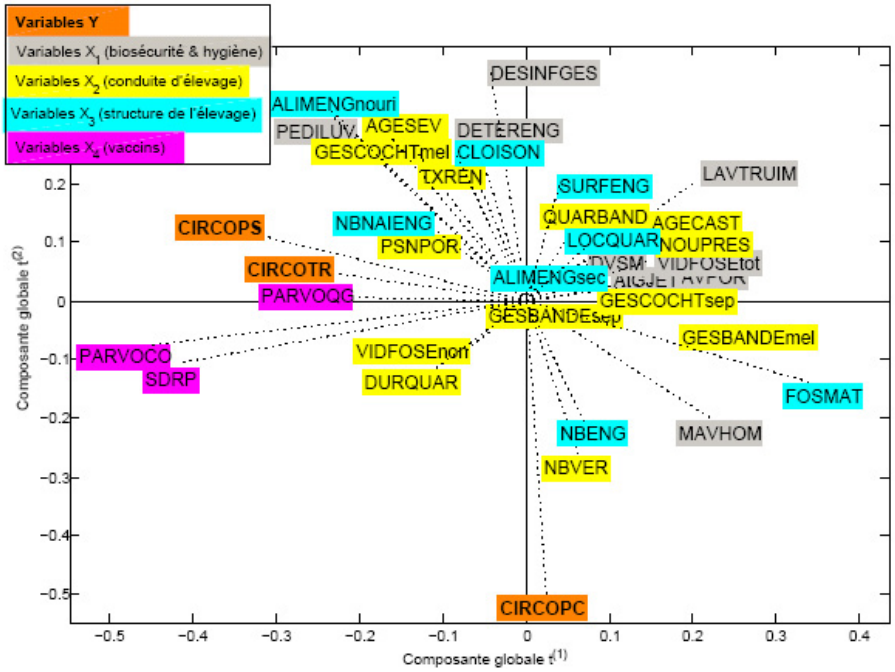
relatives au tableau X_4 résumant les co-facteurs infectieux et vaccins. La méthode *ACPVI* multibloc itérative explique aussi la variable *CIRCOPC* opposée cette fois-ci à *CIRCOPS*. Il faut noter que la carte factorielle de cette dernière méthode positionne de façon particulière les variables de chaque tableau X_k : les variables du tableau X_2, et dans une moindre mesure X_4, sont positionnées sur l'axe opposant *CIRCOPC* à *CIRCOPS*, et sont orthogonales à l'axe sur lequel sont positionnées les variables des tableaux X_3 et X_1.

Comme les quatre cartes factorielles positionnent différemment les variables, seule la carte de l'*ACPVI* multibloc est interprétée ici. Pour déterminer les facteurs de risque relatifs à la proportion d'animaux séropositifs après infection par le circovirus *PCV*2, il est essentiel de raisonner sur l'ensemble des variables Y. En effet, un élevage ayant un profil à risque est un élevage où la proportion de truies et de porcelets séropositifs est faible (la truie transmet peu d'anticorps à ses porcelets par le colostrum) et où en revanche la proportion de porcs à l'engrais séropositifs est élevée (forte pression d'infection virale) [Rose *et al.*, 2003a]. Sur la carte factorielle de l'*ACPVI* multibloc, les variables ayant des coordonnées négativement corrélées à la composante $t^{(1)}$ sont donc associées au profil d'élevage à risque. La variable *FOSMAT*, relative à la profondeur des préfosses en maternité, par exemple, est interprétée comme un facteur de risque pour l'élevage. A l'inverse, les variables ayant des coordonnées positivement corrélées à la composante $t^{(1)}$ sont associées à un profil d'élevage présentant peu de risque. La variable *PEDILUV*, définissant

l'utilisation d'un pédiluve dans chaque salle de l'élevage, est interprétée comme un facteur de risque à effet protecteur pour l'élevage.

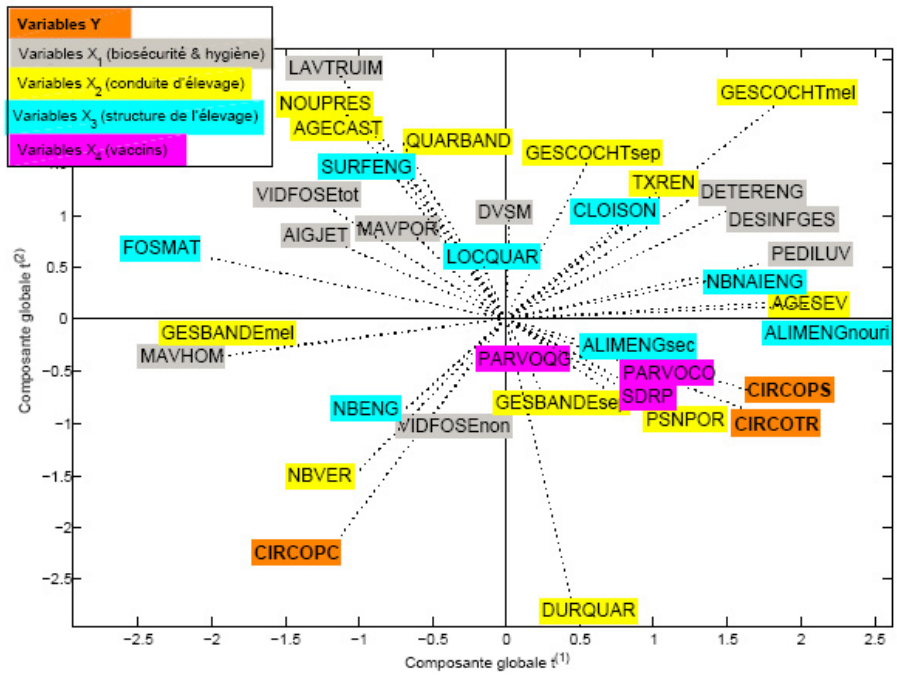


(a) *LRR* multibloc

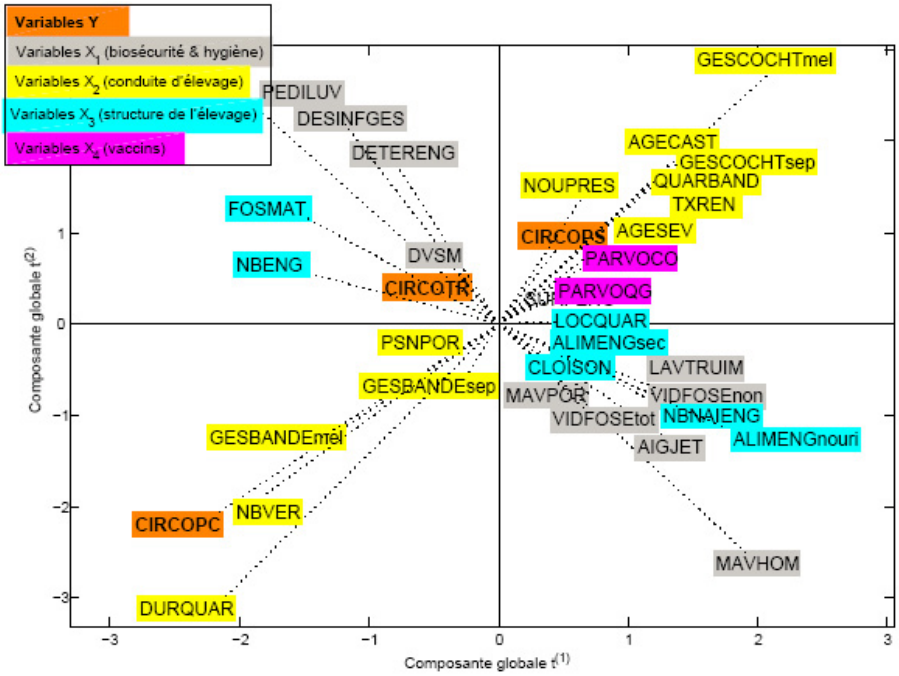


(b) *PLS* multibloc

FIG. 8.6 – Représentation factorielle de l'ensemble des variables sur le plan des composantes globales $(t^{(1)}, t^{(2)})$.



(a) *ACPVI* multibloc



(b) *ACPVI* multibloc à résolution itérative

FIG. 8.7 – Représentation factorielle de l'ensemble des variables sur le plan des composantes globales $(t^{(1)}, t^{(2)})$.

Interprétation du plan des individus

Il est intéressant de mettre en regard des cartes factorielles des variables, les plans des individus, pour les quatre méthodes étudiées (figure 8.8). La couleur des individus est liée à la variable à expliquer *CIRCOPC*, qui représente la proportion de porcs charcutiers séropositifs après infection par le circovirus *PCV*2. Les individus colorés en jaune sont ceux pour lesquels la variables *CIRCOPC* a de faibles valeurs, et ceux qui sont colorés en rouge sont ceux qui ont des valeurs plus élevées pour cette même variable. L'évolution de cette coloration est conforme à la position de la variable *CIRCOPC* sur les plans des composantes $(t^{(1)}, t^{(2)})$ (figures 8.6 et 8.7).

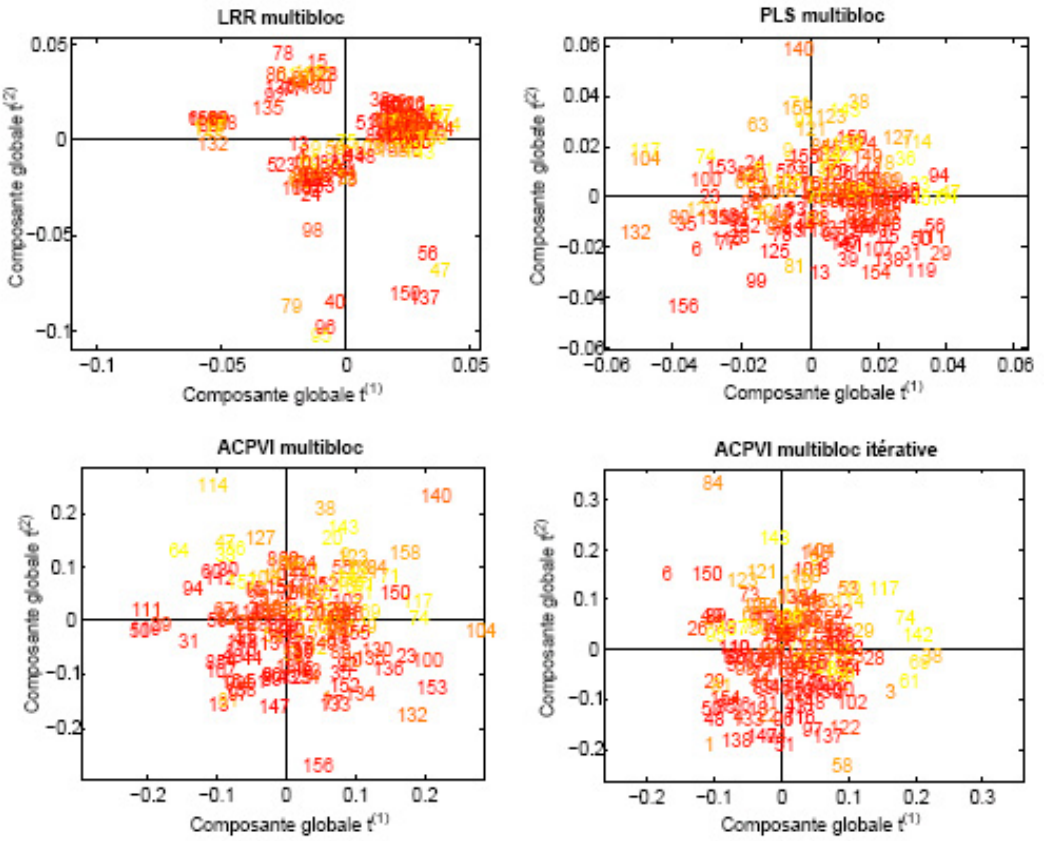


FIG. 8.8 – Représentation factorielle des individus sur le plan des composantes globales $(t^{(1)}, t^{(2)})$.

8.3 Prédiction à partir de tableaux structurés en blocs

8.3.1 Evolution de la norme du vecteur de coefficients

La norme du vecteur de coefficients de régression du modèle où une seule composante $t^{(1)}$ est retenue, $\|\beta^{(1)}\|$, varie en fonction des valeurs des paramètres des deux continuums étudiés (figure 8.9). Le continuum *LRR* multibloc donne des

valeurs de $\|\beta^{(1)}\|$ plus élevées pour les méthodes les plus orientées vers l'explication des variables Y ; il s'agit de la méthode *PLS* multibloc en comparaison à la méthode *ACOM*. On note que le continuum $ACPVI-PLS$ multibloc donne des résultats comparables au continuum dont il est issu dans le cas de deux tableaux (paragraphe 5.3.1 page 83).

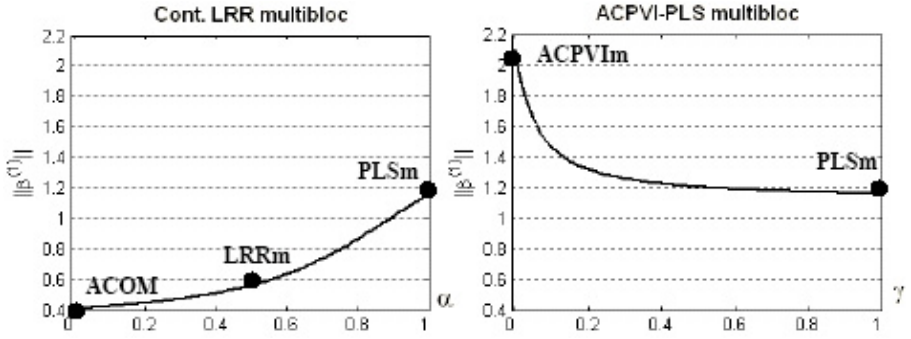


FIG. 8.9 – Comparaison de la norme du vecteur de coefficients $\beta^{(1)}$ pour les continuums *LRR* multibloc et $ACPVI-PLS$ multibloc. Les cas particuliers de ces continuums sont aussi indiqués.

8.3.2 Nombre optimal de dimensions

Résultats pour les méthodes $(K+1)-$**tableaux**

La figure 8.10 illustre l'évolution de l'erreur moyenne de calibration et de validation selon le nombre de composantes conservées dans le modèle, pour les quatre méthodes étudiées. L'erreur moyenne de calibration ($RMSE_C$) illustre la qualité de l'ajustement du modèle aux données. L'erreur moyenne de validation ($RMSE_V$) illustre la qualité prédictive du modèle. Cette dernière est calculée grâce à la procédure validation croisée décrite dans le paragraphe 3.2.2 page 55 sur la base de ($m = 500$) simulations. En terme d'ajustement du modèle aux données, l'*ACPVI* multibloc puis la méthode *PLS* multibloc sont les plus performantes sur ce jeu de données. On retrouve les résultats donnés par la figure 8.3, illustrant l'inertie du tableau Y expliquée par les composantes t. En effet, l'évolution de cette inertie est une autre façon de mesurer la qualité de l'ajustement du modèle aux données. La qualité prédictive générale du modèle est la meilleure pour la méthode *LRR* multibloc. Les méthodes *PLS* multibloc et *ACPVI* multibloc sont les moins prédictives sur cet exemple. Il faut noter que le modèle prédictif optimal (minimisation de l'erreur moyenne de validation) est donné par la *LRR* multibloc ou l'*ACPVI* multibloc à résolution itérative associée à une composante ($RMSE_V = 0.0808$).

Résultats pour les continuums $(K+1)-$**tableaux**

Les figures 8.11 et 8.12 illustrent l'évolution de l'erreur moyenne de calibration et de validation selon le nombre de composantes conservées dans le modèle, pour les deux continuums étudiés. L'erreur moyenne de validation ($RMSE_V$) est calculée

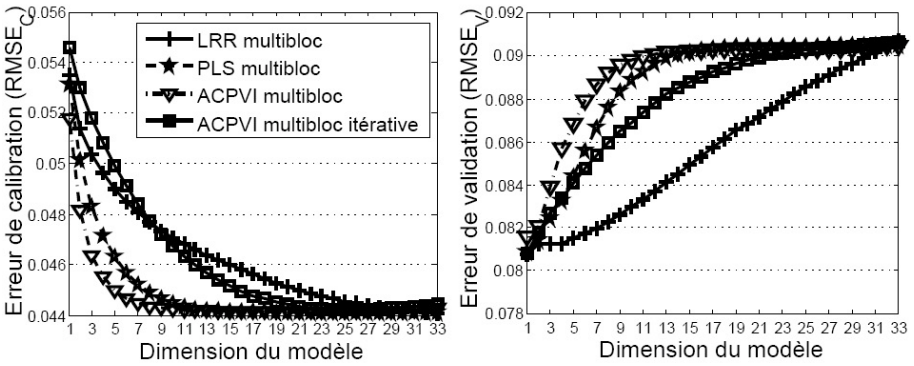


FIG. 8.10 – Erreur moyenne de calibration ($RMSE_C$) et de validation ($RMSE_V$) pour les méthodes *LRR* multibloc, *PLS* multibloc, *ACPVI* multibloc et *ACPVI* multibloc itérative.

grâce à la procédure validation croisée sur la base de ($m = 200$) simulations. Les paramètres α et γ_1 associés aux continuums varient entre 0 et 1 avec un pas de 0.01. Comme les paramètres des continuums sont déterminés de façon à minimiser l'erreur de prédiction, les continuums donnent des résultats optimaux pour la prédiction des variables Y par l'ensemble des variables X.

La figure 8.11 montre que, sur cet exemple, le continuum *LRR* multibloc a une moins bonne qualité de d'ajustement du modèle aux données que les méthodes *LRR* multibloc et surtout que *PLS* multibloc. Par contre, ce continuum apporte une légère amélioration par rapport à la méthode *ACOM* en terme de qualité de prédiction. En effet, les valeurs moyennes des paramètres α optimaux sont inférieures à 0.4, c'est à dire proches des valeurs prises pour la méthode *ACOM* ($\alpha = 0$). Il faut noter que les résultats présentés ne sont pas exactement ceux de l'*ACOM* (sauf pour la solution d'ordre un) car les déflations des tableaux sont réalisées sur les composantes globales et non pas sur les axes partiels comme c'est le cas pour la méthode originelle.

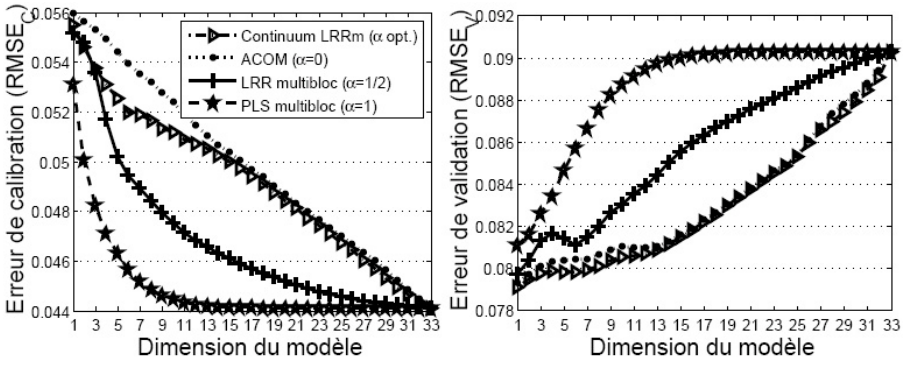


FIG. 8.11 – Erreur moyenne de calibration ($RMSE_C$) et de validation ($RMSE_V$) pour le continuum *LRR* multibloc (α optimum) et ses cas particuliers : *ACOM* ($\alpha = 0$), *LRR* multibloc ($\alpha = 1/2$) et *PLS* multibloc ($\alpha = 1$).

Sur cet exemple, peu de différences de résultats apparaissent entre les méthodes

ACPVI multibloc et *PLS* multibloc, ainsi qu'avec le continuum situé entre celles-ci (figure 8.12). L'*ACPVI* multibloc donne des résultats apparaissant tout de même comme un peu mieux ajustés aux données, mais prédisant légèrement moins bien Y à partir de l'ensemble des variables X. Le continuum donne des résultats comparables à ceux de la méthode *PLS* multibloc. Les valeurs moyennes des paramètres γ_1 optimum sont de l'ordre de $0.6-0.7$ (et donc proche de la *PLS* multibloc, cas particulier associé à $\gamma_1 = 1$) du moins sur les 15 premières dimensions.

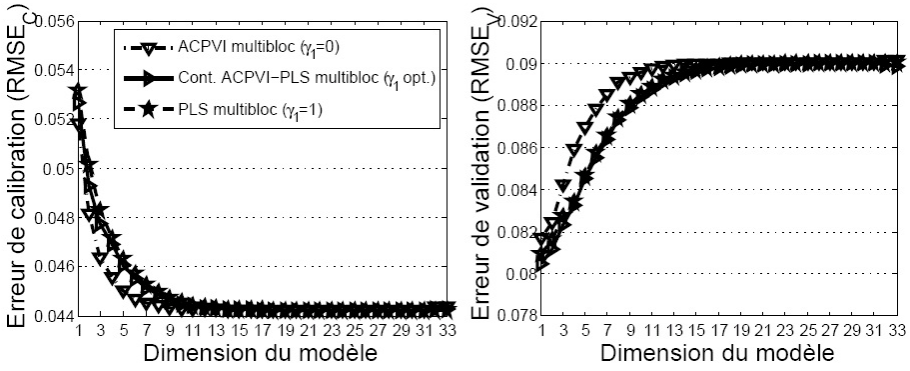


FIG. 8.12 – Erreur moyenne de calibration ($RMSE_C$) et de validation ($RMSE_V$) pour le continuum $ACPVI-PLS$ multibloc (γ_1 optimum) et ses cas particuliers : *ACPVI* multibloc ($\gamma_1 = 0$) et *PLS* multibloc ($\gamma_1 = 1$).

Sélection du nombre optimal de dimensions pour les méthodes $(K+1)-$tableaux

Les indices $Q^{(h)2}$ et $Q^{(h)2}_{cum}$ décrits dans le paragraphe 6.2.3 page 109 permettent de faire un compromis entre les résultats donnés par les erreurs moyennes de calibration et de validation. L'apport d'une composante $t^{(h)}$ dans le modèle peut être évalué par la valeur de l'indice $Q^{(h)2}$ (significatif au-delà du seuil empirique de 0.0975). L'apport des composantes $(t^{(1)},\ldots,t^{(h)})$ est évalué par la valeur de l'indice $Q^{(h)2}_{cum}$ (significatif au-delà du seuil empirique de 0.5). Les résultats, pour les quatre méthodes comparées, sont donnés dans le tableau 8.3. Ces indices donnent des résultats proches et ne permettent pas de bien différencier les méthodes. Pour l'ensemble des méthodes étudiées, seul l'apport de la première composante est significatif. Trois ou quatre composantes sont nécessaires pour prédire correctement le tableau Y.

8.3.3 Influence des blocs et des variables dans l'explication de Y

Influence des blocs dans l'explication du tableau Y

Les résultats du paragraphe précédent sont utilisés pour choisir le nombre optimal de dimensions du modèle. Comme l'objectif est d'avoir une méthode à la fois descriptive et prédictive, l'indice $Q^{(h)2}_{cum}$ est utilisé pour sélectionner le modèle optimal. Un modèle à $(h = 4)$ dimensions est donc sélectionné pour les quatre méthodes étudiées. L'importance des blocs dans l'explication des variables Y est donnée dans

Méthode	**Dimension** h	$Q^{(h)2}$	$Q^{(h)2}_{cum}$
LRR multibloc	1	0.9803 (*)	0.9803 (*)
	2	-1.3082 (NS)	0.0.9545 (*)
	3	-1.501 (NS)	0.8862 (*)
	4	-1.5998 (NS)	0.7041 (*)
	5	-1.6984 (NS)	0.2016 (NS)
PLS multibloc	1	0.9802 (*)	0.9802 (*)
	2	-1.3492 (NS)	0.0.9536 (*)
	3	-1.7053 (NS)	0.8744 (*)
	4	-1.9713 (NS)	0.6269 (*)
	5	-2.2041 (NS)	-0.1956 (NS)
ACPVI multibloc	1	0.9799 (*)	0.9799 (*)
	2	-1.513 (NS)	0.9495 (*)
	3	-2.0353 (NS)	0.8468 (*)
	4	-2.4161 (NS)	0.4165 (NS)
	5	-2.6353 (NS)	-0.903 (NS)
ACPVI mult.iter.	1	0.9803 (*)	0.9803 (*)
	2	-1.2462 (NS)	0.9557 (*)
	3	-1.4358 (NS)	0.8921 (*)
	4	-1.5897 (NS)	0.7206 (*)
	5	-1.7384 (NS)	0.2349 (NS)

TAB. 8.3 – Valeurs des indices $Q^{(h)2}$ et $Q^{(h)2}_{cum}$ selon les cinq premières dimensions du modèle pour les méthodes *LRR* multibloc, *PLS* multibloc, *ACPVI* multibloc et *ACPVI* multibloc itérative. L'astérisque indique un apport significatif, *NS* indique un apport non significatif.

le tableau 8.4 grâce à la moyenne des coefficients a_k^2 sur les quatre premières dimensions (paragraphe 6.2.2 page 107).

La méthode *LRR* multibloc donne plus de poids aux tableaux X_3 (structure de l'élevage) et X_4 (co-facteurs infectieux et vaccins) que les autres méthodes. A l'inverse, les méthodes *ACPVI* multibloc à résolution directe et itérative donnent plus de poids aux tableaux X_1 (biosécurité et hygiène) et X_2 (conduite d'élevage). La méthode *PLS* multibloc donne un poids relativement équilibré aux quatre tableaux dans l'explication de Y.

Influence des variables X dans l'explication du tableau Y

L'influence des variables X, issues des différents tableaux X_k ($k = 1, \ldots, 4$) est donnée par les coefficients de régression des variables X pour expliquer les variables Y. Afin de ne pas alourdir l'interprétation des résultats, seuls les coefficients de régression de la méthode *LRR* multibloc pour un modèle optimal à quatre dimensions sont donnés. En effet, sur cet exemple, cette méthode fournit un bon compromis entre la description et l'explication des données. Les écarts types et intervalles de variabilité à 95% sont calculés à partir des résultats issus des ($m = 500$) simulations.

L'interprétation de ces coefficients de régression est faite au travers des *odds ratio*

Méthode	**Bloc** X_1	**Bloc** X_2	**Bloc** X_3	**Bloc** X_4
LRR multibloc ($h = 4$ *dim.*)	14.9%	11.1%	**43.5%**	**30.5%**
PLS multibloc ($h = 4$ *dim.*)	27.3%	22.7%	20.1%	29.9%
ACPVI multibloc ($h = 4$ *dim.*)	**30.2%**	**33.3%**	19.8%	16.7%
ACPVI mult.iter. ($h = 4$ *dim.*)	**30.8%**	**35.5%**	22.1%	11.6%

TAB. 8.4 – Importance des blocs X_k ($k = 1,\ldots,4$) dans le modèle liant X à Y pour les méthodes *LRR* multibloc, *PLS* multibloc, *ACPVI* multibloc et *ACPVI* multibloc itérative. X_1=biosécurité & hygiène, X_2=conduite d'élevage, X_3=structure de l'élevage et X_4=co-facteurs infectieux & vaccins.

et de leurs intervalles de variabilité (définis paragraphe 1.2.3 page 25), comme c'est l'usage pour les données d'épidémiologie animale. L'interprétation des facteurs de risque de l'ensemble de l'élevage est basé sur l'interprétation conjointe des *odds ratio* liés aux trois variables Y à expliquer. Un élevage au profil protecteur vis à vis de la pression d'infection par le circovirus *PCV*2 est un élevage où la proportion de porcelets et de truies présentant une séroconversion après infection par le circovirus *PCV*2 est élevée et où la proportion de porcs à l'engrais présentant une séroconversion est faible. Un facteur de risque à effet protecteur est donc une variable explicative pour laquelle l'*odds ratio* est supérieur à un (valeur un non comprise dans l'intervalle de variabilité) pour les variables à expliquer *CIRCOPS* et *CIRCOTR*, et est inférieur à un pour la variable à expliquer *CIRCOPC*. A l'inverse, un facteur de risque est une variable explicative pour laquelle l'*odds ratio* est inférieur à un pour les variables à expliquer *CIRCOPS* et *CIRCOTR* et est supérieur à un pour la variable à expliquer *CIRCOPC*. Une variable explicative dont les *odds ratio* contiennent la valeur un dans ses intervalles de variabilité n'est pas considérée comme ayant une influence significative sur les variables à expliquer.

L'épidémiologiste utilise le sens et la significativité des *odds ratio* pour orienter des actions à mener en vue d'améliorer le statut sanitaire du troupeau vis à vis de la pression d'infection par le circovirus *PCV*2. Afin de clarifier l'interprétation, nous donnons tout d'abord le profil de l'élevage pouvant permettre d'améliorer la situation chez le porc à l'engrais. Il consiste à :

1. utiliser un pédiluve dans chaque salle de l'élevage,
2. désinfecter les salles de gestation,
3. mélanger les cochettes (=truies nullipares) et les truies adultes,

Bloc	Variables X	CIRCOPS	CIRCOPC	CIRCOTR
	MAVPOR	-0,01 [-0,23 ;0,21]	-0,03 [-0,27 ;0,20]	-0,01 [-0,23 ;0,20]
	MAVHOM	-0.16 [-0,41 ;0,10]	0.24 [-0,03 ;0,50]	-0.09 [-0,33 ;0,16]
	PEDILUV	0.23 [0,00 ;0,47]	**-0.42** [-0,69 ;-0,14]	0.15 [-0,05 ;0,35]
	AIGJET	-0.06 [-0,31 ;0,19]	0.10 [-0,17 ;0,37]	-0.02 [-0,28 ;0,23]
X_1	DETERENG	0.11 [-0,16 ;0,37]	-0.26 [-0,54 ;0,02]	0.10 [-0,18 ;0,38]
	VIDFOSEnon	-0.01 [-0,25 ;0,23]	0.09 [-0,19 ;0,36]	-0.03 [-0,28 ;0,21]
	VIDFOSEtot	0.00 [-0,22 ;0,22]	-0.09 [-0,33 ;0,45]	-0.03 [-0,27 ;-0,24]
	DESINFGES	0.15 [-0,08 ;0,39]	**-0.36** [-0,62 ;-0,11]	0.12 [-0,15 ;0,38]
	LAVTRUIM	0.00 [-0,25 ;0,25]	-0.08 [-0,36 ;0,20]	-0.01 [-0,24 ;0,23]
	DVSM	0.00 [-0,21 ;0,20]	-0.08 [-0,36 ;0,20]	-0.01 [-0,22 ;0,19]
	GESBANDEmel	-0.11 [-0,35 ;0,14]	0.15 [-0,09 ;0,39]	-0.08 [-0,36 ;0,19]
	GESBANDEsep	0.07 [-0.14 ;0.29]	-0.04 [-0,29 ;0,21]	0.03 [-0,18 ;0,23]
	GESCOCHTmel	0.17 [-0,05 ;0,40]	**-0.31** [-0,57 ;-0,06]	0.12 [-0,11 ;0,35]
	GESCOCHTsep	-0.11 [-0,36 ;0,13]	0.10 [-0,16 ;0,35]	-0.06 [-0,30 ;0,19]
	NOUPRES	0.00 [-0,25 ;0,25]	-0.02 [-0,29 ;0,25]	0.00 [-0,25 ;0,25]
X_2	AGECAST	0.00 [-0,22 ;0,23]	-0.14 [-0,36 ;0,08]	0.01 [-0,23 ;0,24]
	TXREN	0.08 [-0,16 ;0,32]	-0.20 [-0,48 ;0,07]	0.07 [-0,18 ;0,31]
	AGESEV	0.19 [0,03 ;0,41]	**-0.36** [-0,57 ;-0,15]	0.11 [-0,12 ;0,33]
	PSNPOR	0.04 [-0,19 ;0,27]	-0.13 [-0,35 ;0,09]	0.04 [-0,20 ;0,27]
	QUARBAND	0.03 [-0,21 ;0,27]	-0.10 [-0,35 ;0,15]	0.01 [-0,23 ;0,26]
	DURQUAR	-0.01 [-0,24 ;0,21]	0.13 [-0,13 ;0,39]	-0.02 [-0,25 ;0,21]
	NBVER	-0.13 [-0,33 ;0,07]	**0.29** [0,07 ;0,51]	0.08 [-0,29 ;0,12]
	NBENG	-0.08 [-0,30 ;0,14]	**0.32** [0,08 ;0,55]	-0.07 [-0,34 ;0,19]
	NBNAIENG	0.10 [-0,11 ;0,30]	-0.08 [-0,33 ;0,17]	0.06 [-0,15 ;0,27]
	CLOISON	0.19 [-0,01 ;0,39]	**-0.32** [-0,54 ;-0,09]	0.11 [-0,11 ;0,34]
X_3	ALIMENGnourri	**0.29** [0,04 ;0,53]	**-0.52** [-0,79 ;-0,24]	0.17 [-0,03 ;0,37]
	ALIMENGsec	-0.02 [-0,23 ;0,19]	-0.05 [-0,28 ;0,18]	-0.02 [-0,25 ;0,20]
	SURFENG	0.11 [-0,19 ;0,41]	**-0.37** [-0,63 ;-0,11]	0.08 [-0,19 ;0,35]
	LOCQUAR	0.02 [-0,19 ;0,24]	-0.16 [-0,39 ;0,08]	0.05 [-0,18 ;0,28]
	FOSMAT	-0.16 [-0,44 ;0,11]	0.19 [-0,11 ;0,50]	-0.10 [-0,36 ;0,16]
	SDRP	0.24 [-0,01 ;0,50]	-0.10 [-0,34 ;0,14]	**0.26** [0,02 ;0,50]
X_4	PARVOQG	**0.28** [0,02 ;0,54]	-0.01 [-0,28 ;0,26]	-0.03 [-0,28 ;0,21]
	PARVOCO	**0.42** [0,16 ;0,68]	0.12 [-0,13 ;0,40]	0.15 [-0,09 ;0,39]

TAB. 8.5 – Coefficients de régression et leurs intervalles de variabilité à 95%, du modèle liant X à Y, pour la méthode *LRR* multibloc avec ($h = 4$) dimensions.

4. retarder l'âge au sevrage des porcelets,
5. limiter le nombre de verrats introduits dans l'élevage,
6. limiter le nombre d'élevages engraisseurs dans un rayon de 2 km,
7. mettre des cloisons entre les préfosses en engraissement,
8. préférer l'alimentation *nourrisoupe* plutôt que *soupe*,
9. augmenter la surface des cases en engraissement.

Le profil d'élevage pouvant permettre d'améliorer la situation chez les truies et les porcelets consiste à :

1. préférer l'alimentation *nourrisoupe* plutôt que *soupe* lors de l'engraissement,

2. vacciner les truies contre le virus *SDRP*,
3. utiliser le même antigène contre le parvovirus en quarantaine et lors de la gestation des cochettes et truies,
4. augmenter la proportion de cochettes ayant réagi positivement à l'infection au parvovirus.

Bloc	Variables X	CIRCOPS	CIRCOPC	CIRCOTR
	MAVPOR	0,99 [0,80 ;1,23]	0,97 [0,76 ;1,23]	0,99 [0,80 ;1,23]
	MAVHOM	0,85 [0,66 ;1,10]	1,27 [0,97 ;1,65]	0,92 [0,72 ;1,17]
	PEDILUV	1,26 [1,00 ;1,61]	**0,66** [0,50 ;0,87]	1,16 [0,96 ;1,42]
	AIGJET	0,94 [0,73 ;1,21]	1,10 [0,84 ;1,45]	0,98 [0,76 ;1,26]
X_1	DETERENG	1,11 [0,85 ;1,45]	0,77 [0,58 ;1,02]	1,11 [0,84 ;1,46]
	VIDFOSEnon	0,99 [0,78 ;1,26]	1,09 [0,83 ;1,44]	0,97 [0,76 ;1,23]
	VIDFOSEtot	1,00 [0,81 ;1,25]	0,92 [0,72 ;1,16]	0,99 [0,76 ;1,27]
	DESINFGES	1,16 [0,92 ;1,47]	**0,70** [0,54 ;0,90]	1,12 [0,86 ;1,46]
	LAVTRUIM	1,00 [0,78 ;1,28]	0,92 [0,70 ;1,22]	0,99 [0,78 ;1,26]
	DVSM	1,00 [0,81 ;1,23]	0,92 [0,70 ;1,22]	0,99 [0,80 ;1,21]
	GESBANDEmel	0,90 [0,70 ;1,15]	1,16 [0,91 ;1,48]	0,92 [0,70 ;1,21]
	GESBANDEsep	1,07 [0,87 ;1,33]	0,96 [0,75 ;1,23]	1,03 [0,84 ;1,26]
	GESCOCHTmel	1,19 [0,95 ;1,49]	**0,73** [0,57 ;0,94]	1,12 [0,89 ;1,42]
	GESCOCHTsep	0,89 [0,70 ;1,14]	1,10 [0,85 ;1,42]	0,95 [0,74 ;1,21]
	NOUPRES	1,00 [0,78 ;1,28]	0,98 [0,75 ;1,28]	1,00 [0,78 ;1,28]
X_2	AGECAST	1,00 [0,80 ;1,26]	0,87 [0,70 ;1,09]	1,01 [0,80 ;1,27]
	TXREN	1,08 [0,85 ;1,37]	0,82 [0,62 ;1,08]	1,07 [0,84 ;1,37]
	AGESEV	1,21 [0,97 ;1,51]	**0,70** [0,57 ;0,86]	1,11 [0,89 ;1,39]
	PSNPOR	1,04 [0,83 ;1,31]	0,88 [0,70 ;1,10]	1,04 [0,82 ;1,32]
	QUARBAND	1,03 [0,81 ;1,31]	0,91 [0,71 ;1,16]	1,01 [0,79 ;1,29]
	DURQUAR	0,99 [0,79 ;1,24]	1,14 [0,88 ;1,48]	0,98 [0,78 ;1,23]
	NBVER	0,88 [0,72 ;1,07]	**1,34** [1,08 ;1,67]	0,92 [0,75 ;1,13]
	NBENG	0,92 [0,74 ;1,15]	**1,37** [1,08 ;1,74]	0,93 [0,71 ;1,21]
	NBNAIENG	1,10 [0,90 ;1,35]	0,92 [0,72 ;1,19]	1,07 [0,86 ;1,31]
	CLOISON	1,21 [0,99 ;1,48]	**0,73** [0,58 ;0,91]	1,12 [0,89 ;1,40]
X_3	ALIMENGnourri	**1,33** [1,04 ;1,70]	**0,60** [0,45 ;0,79]	1,18 [0,97 ;1,44]
	ALIMENGsec	0,98 [0,80 ;1,21]	0,95 [0,75 ;1,20]	0,98 [0,78 ;1,22]
	SURFENG	1,12 [0,83 ;1,50]	**0,69** [0,53 ;0,90]	1,09 [0,83 ;1,43]
	LOCQUAR	1,02 [0,83 ;1,27]	0,86 [0,68 ;1,08]	1,05 [0,84 ;1,32]
	FOSMAT	0,85 [0,64 ;1,12]	1,21 [0,90 ;1,65]	0,91 [0,70 ;1,17]
	SDRP	1,28 [0,99 ;1,64]	0,90 [0,71 ;1,15]	**1,30** [1,02 ;1,66]
X_4	PARVOQG	**1,32** [1,02 ;1,72]	0,99 [0,76 ;1,30]	0,97 [0,76 ;1,24]
	PARVOCO	**1,52** [1,17 ;1,98]	1,13 [0,88 ;1,50]	1,16 [0,91 ;1,48]

TAB. 8.6 – *Odds ratio* et leurs intervalles de variabilité à 95%, du modèle liant X à Y, pour la méthode *LRR* multibloc avec ($h = 4$) dimensions. Les *OR* significatifs sont en gras.

Conclusion et perspectives

Ce travail de recherche porte sur les méthodes factorielles permettant l'analyse simultanée de plusieurs tableaux. Les contraintes associées au traitement statistique des données d'épidémiologie animale, détaillées dans la partie I, ont amené à centrer ce travail de recherche sur les méthodes d'analyse factorielle qui permettent l'analyse simultanée de plusieurs tableaux, et plus particulièrement sur les méthodes de régression multibloc, qui orientent la description de plusieurs tableaux vers l'explication d'un autre. La prise en compte de la multicolinéarité au sein du tableau des variables explicatives est un point qui est plus particulièrement étudié. De nouveaux outils permettant de contourner ce problème sont proposés, dans le cadre du traitement de deux puis de $(K+1)$ tableaux. Un premier choix est fait de décrire et proposer des méthodes ainsi que des continuums sur la base de critères à maximiser, associés à des contraintes, clairs et conformes aux objectifs définis. Les critères proposés font le lien entre les différents tableaux, au travers des liens entre les composantes associées à chacun de ces tableaux. Un deuxième choix est de présenter principalement des méthodes factorielles. L'ensemble des méthodes décrites dans les revues scientifiques et pouvant s'appliquer au traitement des données d'épidémiologie qui ne répondent pas à ce critère ne sont donc pas décrites dans ce travail de recherche. Dans le cadre des régressions multiblocs proposées, nous avons choisi de donner un rôle central aux composantes globales associées au tableau qui regroupe l'ensemble des variables explicatives. Le choix de réaliser des déflations sur ces composantes globales permet des représentations factorielles de l'ensemble des variables sur un même graphique, ainsi qu'une modélisation prenant en compte l'ensemble des variables explicatives.

La partie II détaille dans un premier temps les méthodes permettant de lier deux tableaux. Les méthodes *ACPVI*, *ACP* associée à une *PCR*, régression *PLS* et analyse canonique sont tout d'abord présentées de façon uniformisée à la fois du point du vue du critère que des contraintes de normes et de déflation. Dans ce cadre, une version modifiée de la *latent root regression*, basée sur la maximisation et la résolution directe d'un critère est proposée. Deux nouveaux continuums, dont les objectifs de traitement sont adaptés aux données d'épidémiologie, sont proposés et comparés aux continuums déjà décrits dans les revues scientifiques. Le premier relie *PCR*, *LRR* et *PLS*, et le second *ACPVI* et *PLS*. L'écriture des critères associés aux méthodes sous forme de continuum permet de démontrer certaines propriétés relatives aux méthodes elles-mêmes. La partie III étend les méthodes précédentes, ainsi que les continuums qui les relient, au cas du traitement de $(K+1)$ tableaux. Plusieurs

méthodes de régression multibloc, adaptées à l'analyse des données d'épidémiologie animale, sont proposées. Tout d'abord une analyse canonique généralisée sous contrainte, dont le mode de résolution illustre sa plus grande robustesse à la multicolinéarité que l'analyse canonique généralisée avec tableau de référence proposée par Kissita [2003]. Puis deux méthodes s'apparentant à des extensions multiblocs de l'*ACPVI* sont développées. Une dernière méthode est proposée, dérivant de l'extension multibloc de la *latent root regression*, dont la robustesse à la multicolinéarité devrait être comparable à celle de la régression *PLS* multibloc.

L'ensemble des méthodes et continuums présentés dans ce travail de recherche a été appliqué à de nombreux jeux de données d'épidémiologie animale. Tous ces résultats ne sont bien entendu pas présentés. Afin de clarifier les chapitres d'application 5 et 8, seuls les résultats des méthodes et des continuums les plus intéressants, ainsi que leurs applications à deux jeux de données seulement, sont présentés. Ce travail de recherche a permis d'appliquer les méthodes multiblocs à un domaine dans lequel elles n'avaient pas encore été appliquées. Ces applications ont suscité un grand intérêt de la part des praticiens. Aucune difficulté d'interprétation n'a été rencontrée par les épidémiologistes et les résultats sont conformes et plus riches que ceux obtenus au préalable par des méthodes statistiques plus classiques (référencées dans les publications du tableau 1.1 page 27). Les possibilités graphiques offertes par ces méthodes sont nombreuses et présentent un grand intérêt pour l'interprétation ainsi que pour la communication des résultats par les épidémiologistes. En effet, ces résultats sont communiqués à d'autres scientifiques, ainsi qu'aux éleveurs et vétérinaires concernés par les résultats obtenus ; ils doivent donc être clairs et accessibles à un public peu familier des méthodes d'analyse factorielle.

Les avancées en termes de traitement des données d'épidémiologie animale sont tout d'abord de permettre la modélisation simultanée de plusieurs variables. Ce nouveau mode de traitement permet ainsi de s'affranchir de la réalisation de plusieurs modèles aux conclusions bien souvent différentes, ainsi qu'à la synthèse des variables à expliquer en une seule variable de synthèse, solution qui n'est pas toujours satisfaisante pour l'épidémiologiste qui connaît la complexité du problème qu'il étudie. La seconde avancée apportée par les méthodes factorielles est de permettre la prise en compte d'un plus grand nombre de variables explicatives qu'avec l'utilisation les méthodes de régression plus usuelles. Ceci évite à l'épidémiologiste un tri délicat et peu satisfaisant. Le fait d'utiliser des méthodes factorielles couplées à des modélisations permet de plus d'associer les avantages de ces deux types de méthodes statistiques, trop souvent dissociés dans la pratique. La troisième et plus importante avancée est l'utilisation de méthodes peu vulnérables à la multicolinéarité. Ce point crucial est à la fois pris en compte par l'application de méthodes dont la résolution est peu sensible à la quasi-colinéarité des variables explicatives, ainsi que par l'utilisation conjointe de la régression orthogonalisée. La dernière avancée est de pouvoir désormais prendre en compte la structure en blocs des variables explicatives. Ceci permet d'équilibrer le poids des blocs dans l'explication des variables Y, mais aussi d'apporter de nouvelles réponses aux épidémiologistes comme la mesure de l'influence des blocs dans l'explication de la maladie.

Les résultats obtenus sur la base d'études de cas sont dans l'ensemble conformes aux propriétés théoriques. Les méthodes pour lesquelles le critère est le moins orienté vers les variables à expliquer, à savoir *PCR*, *LRR* et *LRR* multibloc, sont plus explicatives des variables X que des variables Y. Les méthodes *ACPVI* et *ACPVI* multibloc apparaissent toujours comme les méthodes les plus explicatives des variables Y et les moins explicatives des variables X. La régression *PLS* apparaît intermédiaire entre les méthodes *LRR* et *ACPVI* pour le premier jeu de données (chapitre 5) et proche de l'*ACPVI* pour le second jeu de données (chapitre 8). Ces différences ne semblent pas provenir de la structure des données en deux tableaux ou $(K+1)$ tableaux, mais plutôt de la différence en termes de multicolinéarité et de nombre d'individus des deux jeux de données utilisés. Nous pouvons noter que l'utilisation des continuums permettant de lier les méthodes les plus intéressantes, apporte du point de vue pratique assez peu d'amélioration. Le continuum est toujours proche de la méthode existante la plus performante et l'améliore peu en pratique. Les changements ne sont pas graduels comme l'on pourrait s'y attendre. Il conviendrait d'étayer ce point par l'analyse d'autres jeux de données et par des études de simulations.

Toutes ces avancées méthodologiques permettent de mieux prendre en compte la complexité des données d'épidémiologie animale. Cependant, certaines limites apparaissent et devront être résolues par la suite afin de permettre une utilisation plus routinière de ces méthodes dans la détermination des facteurs de risque des maladies ou des problèmes étudiés en sécurité sanitaire des aliments.

Deux problématiques majeures relatives à la structure des données d'épidémiologie animale ne sont pas explorées dans ce travail de recherche. La problématique la plus importante est relative à la structure hiérarchisée des individus, actuellement prise en compte au travers d'effets aléatoires dans les régressions. Cette problématique est sûrement la principale limite des méthodes proposées dans le cadre de ce travail de recherche pour leur application systématique au traitement des données d'épidémiologie animale. En effet, ne pas prendre en compte la structure hiérarchisée des individus rend la modélisation utilisée dans le cadre des régressions multiblocs peu subtile en comparaison à celle utilisée dans les modèles linéaires généralisées à effets mixtes. Actuellement, il est possible d'en faire une analyse descriptive en la visualisant sur les plans factoriels des individus. Lors de la régression orthogonalisée, ces effets aléatoires devraient pouvoir être intégrés. La seconde problématique non traitée est relative aux données manquantes, de l'ordre de 10% en moyenne dans ce type d'enquête pour les données collectées à l'*AFSSA* (paragraphe 1.3.2 page 26). Les solutions habituellement utilisées sont soit la suppression d'individus ou de variables concernés, soit l'utilisation des stratégies d'imputation. Ces solutions ne sont bien entendu pas satisfaisantes ; d'autres solutions plus performantes existent et font l'objet de nombreuses recherches actuellement [Allison, 2002]. L'intégration de l'algorithme *NIPALS*, développé initialement dans le cadre de la régression *PLS*, devrait être possible.

D'autres problématiques plus secondaires pourraient permettre une meilleure adéquation des méthodes statistiques présentées aux spécificités des données d'épidémiologie animale. Les variables qualitatives ont été intégrées dans ce travail de façon très classique grâce à l'utilisation des indicatrices des classes. Un meilleur codage pourrait être proposé en tenant compte du lien entre la variable qualitative et le tableau des variables à expliquer par exemple. Il faut noter de plus que le lien entre les variables explicatives et les variables à expliquer utilisé dans ce travail de recherche est linéaire. Dans le cas où les variables à expliquer sont qualitatives, ce lien linéaire est moins adapté que le lien logistique par exemple. Des améliorations pourraient aussi être apportées à la validation croisée, point crucial de validation du modèle, en intégrant l'information contenue dans les données. Les méthodes dites de validation croisée semi-paramétrique proposent, dans le cas où l'on dispose d'un modèle, de ne pas réaliser la validation sur les données brutes mais plutôt sur les résidus centrés du modèle [Freedman, 1981]. L'utilisation de techniques de validation croisée de type *bootstrap* permettrait de plus de calculer de réels intervalles de confiance (*i.e.* calculés sur N individus) associés aux coefficients de régression du modèle. L'extension des méthodes proposées au cas de plusieurs tableaux Y devrait pouvoir être réalisée sans difficultés majeures. Cette extension pourrait permettre l'étude des facteurs de risque de l'évolution d'une maladie au cours du temps par exemple.

D'un point de vue plus théorique, l'utilisation de données simulées dont la structure est parfaitement connue, sera réalisée par la suite pour étudier le comportement des méthodes présentées à la multicolinéarité des variables explicatives notamment. Ce développement n'a pas été réalisé dans le cadre de ce travail car l'application aux données réelles disponibles a été privilégiée. L'approche *PLS* offre un cadre général à un grand nombre de ces méthodes qu'il conviendrait d'explorer [Wold, 1982; Tenenhaus, 1998, 1999; Tenenhaus *et al.*, 2005a,b]. Les contraintes associées aux méthodes présentées devraient pouvoir être reliées aux différents schémas d'estimation (mode A ou B, ainsi que les modèles centroïde, factoriel ou structurel).

Les méthodes de régression multiblocs présentées dans ce travail de recherche sont directement applicables aux données de même structure provenant d'autres domaines ayant des problématiques de traitement similaires, tels que la chimiométrie, la sensométrie, le marketing ou l'écologie. En effet, des données multicorrélées organisées en $(K+1)$ tableaux sont couramment rencontrées dans la pratique. Du fait de l'augmentation du nombre de variables recueillies et de la complexité des questions posées, notamment dans les domaines biologiques, l'utilisation des méthodes multiblocs devrait progressivement se vulgariser.

Annexe : Liste des publications

Chapitre d'ouvrage à comité de lecture

Bougeard, S., Hanafi, M., Noçairi, H. et Qannari, E. M., **2006**. *Multibloc canonical correlation and redundancy analyses for categorical variables : application to epidemiological data (Chap.* 17*)*. Multiple correspondence analysis and related methods. Edited by M.Greenacre and J.Blasius. Chapman & Hall, 393-404.

Publications dans des revues à comité de lecture

Acceptées

Bougeard, S., Hanafi, M., et Qannari E. M., **2007**. Multiblock latent root regression. Application to epidemiological data. *Computational statistics and data analysis*, 22 : 209-222.

Bougeard, S., Hanafi, M., et Qannari E. M., **2007**. *ACPVI* multibloc. Application à des données d'épidémiologie animale. *Journal de la Société Française de Statistique*, 148(4) : 77-94.

Dory, D., Béven, V., Torché, A. M., Bougeard, S., Cariolet, R. et Jestin, A., **2005**. CpGmotif in *ATCGAT* hexamer improves *DNA*-vaccine efficiency against lethal Pseudorabies virus infection in pigs. *Vaccine*, 23 : 4532-4540.

Gravier, R., Dory, D., Laurentie, M., Bougeard, S., Cariolet, R. et Jestin, A., **2007**. In vivo Tissue Distribution and Kinetics of a Pseudorabies Virus Plasmid *DNA* Vaccine after Intramuscular Injection in Swine. *Vaccine*, 25 : 6930-6938.

Gravier, R., Dory, D., Rodriguez, F., Bougeard, S., Beven, V., Cariolet, R. et Jestin, A., **2007**. Immune and protective abilities of ubiquitinated and non-ubiquitinated pseudorabies virus glycoproteins. *Acta Virologica*, 51 : 35-45.

Guionie, O., Toquin, D., Sellal, E., Bouley, S., Zwingelstein, F., Allée, C., Bougeard, S., Lemière, S. et Eterradossi, N., **2007**. Laboratory evaluation of a quantitative real-time reverse transcription *PCR* assay for the detection and identification of the four subgroups of avian metapneumovirus. *Journal of Virological Methods*, 139 :

150-158.

Le Potier, M.F., Le Dimna, M., Kuntz-Simon, G., Bougeard, S. et Mesplede, A., **2006**. Validation of a real-time *RT – PCR* assay for rapid and specific diagnosis of classical swine fever virus. *Developments in Biologicals*, 126 : 179-186.

Mahé, A., Bougeard, S., Salaün, A., Le Bouquin, S., Pétetin, I., Rouxel, S., Lalande, F., Beloeil, P.A. et Rose, N., *In press*. Bayesian estimation of flock-level sensitivity of detection of Salmonella *Spp.* Enteritidis and Typhimurium according to the sampling procedure in french laying-hen houses. *Preventive Veterinary Medicine*.

Marois, C., Bougeard, S., Gottschalk, M. et Kobisch, M., **2004**. Multiplex *PCR* assay for detection of Straptococcus suis species and serotypes 2 and 1/2 in tonsils of live and dead pigs. *Journal of clinical microbiology*, 42 : 3169-3175.

Soumises pour publication

Bougeard, S., Hanafi, M. et Qannari, M., Continuum redundancy *PLS* regression : a simple continuum approach, application to epidemiological data. *Computational statistics and data analysis*.

Mahé, A., Bougeard, S., Huneau-Salaün, A., Le Bouquin, S., Pétetin, I., Rouxel, S., Lalande, F., Beloeil, P.A. et Rose, N., Estimation de la sensibilité de détection de Salmonella *Spp.* en fonction du nombre de prélèvements dans les élevages de poules pondeuses oeufs de consommation. *Epidémiologie et santé animale*.

Actes de congrès

Bougeard, S., Hanafi, M. et Qannari, E.M., **2006**. Continuum redundancy *PLS* regression : a simple continuum approach. Application to epidemiological data. *Congrès COMPSTAT*, 28 août-1er septembre, Rome (Italie), 657-664.

Bougeard, S., Chauvin, C., Hanafi, M. et Qannari, E.M., **2006**. Description and prediction from multiblock tables, application to epidemiological data. *11th International Symposium on Veterinary Epidemiology and Economics*, 6-11 Août, Cairns (Australie), 447-450.

Bougeard, S., Hanafi, M. et Qannari, E.M., **2006**. ACPVI multibloc, application à des données d'épidémiologie animale. *Congrès de la Société Française de Statistique*, 29 mai-2 juin, Clamart (France).

Bougeard, S., Hanafi, M. et Qannari, E.M., **2005**. Analyse de co-inertie multiple sous contrainte, application à des données d'épidémiologie animale. *Congrès de la*

Société Française de Statistique, 6-10 juin, Pau (France).

Bougeard, S., Hanafi, M. et Qannari, E.M., **2005**. Multiblock latent root regression. *Congrès international PLS and related methods*, 7-9 septembre, Barcelone (Espagne), 141-148.

Bougeard, S., Qannari, E.M. et Noçairi, H., **2003**. Discriminant analysis on categorical variables. *Correspondence Analysis and Related Methods*, 29 juin-2 juillet, Barcelone (Espagne), 10.

Cariolet, R., Oswald, I., Le Digherher, G., Bougeard, S., Cossalter, A.M., Ecobichon, P. et Le Divich, J., **2007**. Aquisition de l'immunité passive chez les porcelets issus de truies exemptes d'organismes pathogènes spécifiques (*EOPS*). 39*iemes journées de la recherche porcine*, 6-8 février, Paris (France), 429-430.

Cariolet, R., Le Diguerher, G., Julou, P., Rose, N., Ecobichon, P., Bougeard, S. et Madec, F., **2004**. Survie et croissance des porcelets au stade maternité dans l'unité *EOPS* de l'*AFSSA* de Ploufragan. 36*iemes journées de la recherche porcine*, 3-5 février, Paris (France), 435-442.

Denis, M., Rose, V., Huneau-Salaün, A., Bougeard, S., Balaine, L. et Salvat, G., **2005**. Les Campylobacters dans les élevages de poulets de chair élevés en plein air. Relation entre les génotypes de Campylobacter et les caractéristiques des élevages. 6*imes journées de la recherche avicole*, 30-31 mars, Saint malo (France).

Guionie, O., Toquin, D., Sellal, E., Bouley, S., Zwingelstein, F., Allée, C., Bougeard, S., Lemière, S. et Eterradossi, N., **2006**. Laboratory evaluation of a quantitative real-time reverse transcription *PCR* assay for the detection and identification of the four subgroups of avian metapneumovirus. *5th International Symposium on avian corona and pneumoviruses and complicating pathogens*, 14-16 mai, Rauischholzhausen (Allemagne).

Le Dimna, M., Kuntz-Simon, G., Louguet, Y., Bougeard, S. et Le Potier, M.F., **2006**. Validation de nouveaux tests de détection par $RT-PCR$ en temps réel du génôme du virus de la peste porcine classique. 38*iemes Journées de la recherche porcine*, 31 janvier-2 février, Paris (France), 365-370.

Qannari, E.M., Noçairi, H., Hanafi, M. et Bougeard, S., **2003**. Multibloc redundancy and *PLS* analyses, application to sensory studies. *Congrès international PLS*, 15-17 septembre, Lisbonne (Portugal).

Rose, N., Bougeard, S., Le Digherher, G., Eveno, E., Jolly, J.P., Larour, G. et Madec, F., **2003**. Influence of different logistic regression analysis on the outcome of a case / control study for post-weaning multisystemic wasting syndrom (*PMWS*) risk factors. *10th Symposium of the International Society for Veterinary Epidemiology and Economics*, 17 – 21 novembre 2003, Vina del Mar (Chili), 105-110.

Posters

Fablet, C., Robinault, C., Jolly, J. P., Dorenlor, V., Eono, F., Eveno, E., Labbé, A., Bougeard, S., Fravalo, P. et Madec, F., **2007**. Estimation of the risk of Salmonella shedding by finishing pigs using a logistic model obtained from a survey. *7th international symposium on the epidemiology and control of foodborne pathogens in pork*, 9-11 mai, Vérone (Italie).

Toquin, D., Guionie, O., Sellal, E., Bouley, S., Allée, C., Zwingelstein, F., Bougeard, S., Cherbonnel, M., Jestin, V. et Eterradossi, N., **2006**. Evaluation d'une *RT – PCR* en temps réel pour la quantification des métapneumovirus aviaires (*AMPV*) et application à l'étude expérimentale de l'excrétion des *AMPV* de sous-groupe *C* chez le canard de barbarie. *VIIIièmes journées francophones de virologie*, 20-21 avril, Paris (France).

Bibliographie

A. Agresti : *Categorical data analysis*. John Wiley and Sons, Hoboken, New Jersey, 2ième édition édition, 2002.

P.D. Allison : *Logistic regression using the SAS system : theory and application*. SAS Institute & Wiley, Cary, NC, USA, 1999.

P.D. Allison : *Missing data*. Sage, Thousand Oaks, 2002.

G. Aumont, D. Gauthier, L. Gruner et G. Matheron : Dynamics of the free-living populations of gastrointestinal trichostrongyles of cattle in a natural grazing system in guadeloupe (french west indies). *Preventive Veterinary Medicine*, 12(3-4):245–258, 1992.

L. Barker et C. Brown : Logistic regression when binary predictor variables are highly correlated. *Statistics in medicine*, 20:1431–1442, 2001.

M. Barker et W. Rayens : Partial least squares for discrimination. *Journal of chemometrics*, 17:166–173, 2003.

J. Barnouin, M. Chassagne et I. Aimo : Dietary factors associated with milk somatic cell counts in dairy cows in Brittany, France. *Preventive Veterinary Medicine*, 21:299–311, 1995.

P. Bastien, V.E. Vinzi et M. Tenenhaus : PLS generalised linear regression. *Computational statistics and data analysis*, 48:17–46, 2005.

F. Beaudeau et C. Fourichon : Estimating relative risk of disease from outputs of logistic regression when the disease is not rare. *Preventive Veterinary Medicine*, 36:243–256, 1998.

P.A. Beloeil, C. Chauvin, K. Proux, F. Madec, P. Fravalo et A. Alioum : Impact of the salmonella status of market-age pigs and the pre-slaughter process on salmonella caecal contamination at slaughter. *Veterinary Research*, 35:513–530, 2004a.

P.A. Beloeil, C. Chauvin, K. Proux, N. Rose, S. Queguiner, E. Eveno, C. Houdayer, V. Rose, P. Fravalo et F. Madec : Longitudinal serological responses to salmonella enterica of growing pigs in a subclinically infected herd. *Preventive Veterinary Medicine*, 60:207–226, 2003.

P.A. Beloeil, P. Fravalo, C. Fablet, J.P. Jolly, E. Eveno, Y. Hascoet, C. Chauvin, G. Salvat et F. Madec : Risk factors for salmonella enterica subsp. enterica shedding by market-age pigs in french farrow-to-finish herds. *Preventive Veterinary Medicine*, 63:103–120, 2004b.

D.A. Belsley, E. Kuh et R.E. Welsch : *Regression diagnostics : identifying influential data and sources of collinearity*. Wiley, Ed., 1980.

J.P. Benzecri : *L'analyse des données . Tome 1 : La taxonomie. Tome 2 : L'analyse des correspondances*. Dunod, Paris, 1973.

R.D. Berghaus, J.E. Lombard, I.A. Gardner et T.B. Farver : Factor analysis of a Johne's disease risk assessment questionnaire with evaluation of factor scores and a subset of original questions as predictors of observed clinical paratuberculosis. *Preventive Veterinary Medicine*, 72(3-4):291–309, 2005.

A. Berglund et S. Wold : A serial extension of multiblock PLS. *Journal of chemometrics*, 13:461–471, 1999.

D. Bertrand, E.M. Qannari et E. Vigneau : Latent root regression analysis : an alternative method to PLS. *Chemometrics and intelligent laboratory systems*, 58:227–234, 2001.

A. Boklund, L. Alban, S. Mortensen et H. Houe : Biosecurity in 116 danish fattening swineherds : descriptive results and factor analysis. *Preventive Veterinary Medicine*, 66(1-4):49–62, 2004.

S. Bougeard, M. Hanafi et E.M. Qannari : Analyse de co-inertie multiple sous contrainte. Application à des données d'épidémiologie animale. *In Congrès de la Société Française de Statistique*, Pau (France), 2005a.

S. Bougeard, M. Hanafi et E.M. Qannari : Multiblock latent root regression. *In Congrès PLS and related methods*, pages 141–148, Barcelone (Espagne), 2005b.

S. Bougeard, M. Hanafi et E.M. Qannari : Multiblock latent root regression. Application to epidemiological data. *Computational statistics and data analysis*, 22 (2):209–222, 2007.

J. Bouyer, D. Hemon, S. Cordier, F. Derriennic, I. Stucker, B. Stengel et J. Clavel : *Epidémiologie. Principes et méthodes quantitatives*. INSERM, 1995.

R. Bro : Multiway calibration. Multilinear PLS. *Journal of chemometrics*, 10:47–61, 1996.

R.J. Brooks et M. Stone : Joint continuum regression for multiple predictants. *Journal of american statistical association*, 89(428):1374–1377, 1994.

P.J Brown et J.V. Zidek : Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1):64–74, 1980.

X. Bry : Estimation empirique d'un modèle à variables latentes comportant des interactions. *Revue de Statistique Appliquée*, 52(3):5–35, 2004.

A.J. Burnham, J.F. Macgregor et R. Viveros : A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of chemometrics*, 13:49–65, 1999.

A.J. Burnham, R. Viveros et J.F. Macgregor : Framework for latent variable multivariate regression. *Journal of chemometrics*, 10:31–45, 1996.

J.D. Carroll : A generalization of canonical correlation analysis to three or more sets of variables. *In 76th annual convention of the American psychological association*, pages 227–228, 1968.

P. Casin : L'analyse discriminante de tableaux évolutifs. *Revue de statistique appliquée*, XLIII(3):73–91, 1995.

P. Casin : L'analyse en composantes principales généralisée. *Revue de statistique appliquée*, XLIV(3):63–81, 1996.

P. Cazes : *Application de l'analyse des données au traitement de problèmes géologiques*. Thèse de doctorat, Faculté des sciences de Paris, 1970.

P. Cazes : Protection de la régression par utilisation de contraintes linéaires et non linéaires. *Revue de statistique appliquée*, XXIII(3):37–57, 1975.

P. Cazes : L'analyse de certains tableaux rectangulaires décomposés en blocs : généralisation de propriétés rencontrées dans l'étude des correspondances multiples. I. Définitions et applications à l'analyse canonique des variables qualitatives. *Les cahiers de l'analyse des données*, 5(2):89–99, 1980.

G. Celeux et J.P. Nakache : *Analyse discriminante sur variables qualitatives*. Polytechnica, Paris, 1994.

C. Chauvin, I. Bouvarel, J.P. Orand, P.A. Beloeil, D. Guillemot et P. Sanders : A pharmaco-epidemiological analysis of factors associated with antimicrobial consumption level in turkey broiler flocks. *Veterinary Research*, 36(2):199–211, 2005.

D. Chessel et M. Hanafi : Analyses de la co-inertie de K nuages de points. *Revue de statistique appliquée*, XLVI(2):35–60, 1996.

D. Chessel et P. Mercier : Couplage de triplets statistiques et liaisons espèces-environnement. *In* LEBRETON J.D. et ASSELAIN B., éditeur : *Biométrie et environnement*, pages 15–44. MASSON, Paris, 1993.

J. Chi, J.A. VanLeeuwen, A. Weersink et G.P. Keefe : Management factors related to seroprevalences to bovine viral-diarrhoea virus, bovine-leukosis virus, mycobacterium avium subspecies paratuberculosis, and neospora caninum in dairy herds in the canadian maritimes. *Preventive Veterinary Medicine*, 55(1):57–68, 2002.

P. Congdon : *Apllied bayesian modelling*. J.Wiley and Sons, Chichester, England, 2003.

P. Congdon : *Bayesian models for categorical data*. J.Wiley and Sons, Chichester, England, 2005.

P.T. Davies et M. K.S. Tso : Procedures for reduced-rank regression. *Appl. Statist.*, 31:244–255, 1982.

B.S. Dayal et J.F. MacGregor : Improved PLS algorithms. *Journal of chemometrics*, 11:73–85, 1997.

T. De Bie, N. Christianini et R. Rosipal : Eigenproblems in pattern recognition. *In* Springer, éditeur : *Handbook of geometric computing : application in pattern recognition, computer vision, neural computing and robotic*. Bayro-Corrochano, E., 2005.

S. De Jong : SIMPLS : an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, 18:251–263, 1993.

S. De Jong : PLS shrinks. *Journal of chemometrics*, 9:323–326, 1995.

S. De Jong et R. W. Farebrother : Extending the relationship between ridge regression and continuum regression. *Chemometrics Intelligent Lab. Systems*, 25:179–181, 1994.

S. De Jong et H.A.L. Kiers : Principal covariates regression. Part I. Theory. *Chemometrics and intelligent laboratory systems*, 14:155–164, 1992.

S. De Jong, B.M. Wise et N.L. Ricker : Canonical partial least squares and continuum power regression. *Journal of chemometrics*, 15:85–100, 2001.

S. Derksen et H.J. Keselman : Backward, forward and stepwise automated subset selection algorithms : frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45:265–282, 1992.

C. Derquenne et C. Hallais : Une méthode alternative à l'approche PLS : comparaison et application aux modèles conceptuels marketing. *Revue de statistique appliquée*, 52(3):37–72, 2004.

I.R. Dohoo, C. Ducrot, C. Fourichon, A. Donald et D. Hurnik : An overview of techniques for dealing with large numbers of independant variables in epidemiologic studies. *Preventive Veterinary Medicine*, 29:221–239, 1997.

I.R. Dohoo, W. Martin et H. Stryhn : *Veterinary epidemiologic research*. Atlantic Veterinary College Inc., University of Prince Edward Island, Prince Edward Island, Canada, 2003.

J.J. Droesbeke, M. Lejeune et G. Saporta : *Modèles statistiques pour données qualitatives*. Technip, Paris, 2005.

L. Duchateau, R.L. Kruska et B.D. Perry : Reducing a spatial database to its effective dimensionality for logistic-regression analysis of incidence of livestock disease. *Preventive Veterinary Medicine*, 32:207–218, 1997.

C. Ducrot et I. Cimarosti : Complémentarité du modèle logistique et de l'analyse des correspondances pour la recherche des facteurs de risque en pathologie animale : application à l'étude des facteurs de risque de l'ecthyma des ovins. *Epidémiologie et santé animale*, 20:126–131, 1991.

H. Erkel-Rousse : Détection de la multicolinéarité dans un modèle linéaire ordinaire : quelques éléments pour un usage averti des indicateurs de Belsley, Kuh et Welsch. *Revue de statistique appliquée*, XLIII(4):19–42, 1995.

A.S. Evans : Causation and disease : a chronological journey. *Am. J. Epidemiol.*, 108:249–258, 1978.

B. Faye et F. Lescourret : Environmental factors associated with lameness in dairy cattle. *Preventive Veterinary Medicine*, 7(4), 1989.

B. Faye, F. Lescourret, N. Dorr, E. Tillard, B. MacDermott et J. McDermott : Interrelationships between herd management practices and udder health status using canonical correspondence analysis. *Preventive Veterinary Medicine*, 32(3-4):171–192, 1997.

R.A. Fisher : The use of multiple measurements in taxonomy problem. *Annals of Eugenics*, 7:179–188, 1936.

I.E. Frank et J.H. Friedman : A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

D.A. Freedman : Bootstrapping regression models. *Annals of Statistics*, 9:1218–28, 1981.

R. Ganaba, M. Bigras-Poulin, D. Bélanger et Y. Couture : Description of cow-calf productivity in Northwestern Quebec and path models for calf mortality and growth. *Preventive Veterinary Medicine*, 24:31–42, 1995.

T.C. Gleason : On redundancy in canonical analysis. *Physiological bulletin*, 83 (6):1004–1006, 1976.

W.J. Goodger, T. Farver, J. Pelletier, P. Johnson, G. De Snayer et J. Galland : The association of milking management practices with bulk tank somatic cell counts. *Preventive Veterinary Medicine*, 15(4):235–251, 1993.

C. Guinot, J. Latreille et Tenenhaus M. : PLS path modeling and analysis of multiple tables. *Chemometrics and Intelligent Laboratory Systems (Special issue on PLS methods)*, 58, 2001.

M. Hanafi : *Structure de l'ensemble des analyses multivariées des tableaux de données à trois entrées : éléments théoriques et appliqués.* Thèse de doctorat, Université Lyon 1, 1997.

M. Hanafi et H.A.L. Kiers : Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Computational statistics and data analysis*, 51(3):1491–1508, 2006.

M. Hanafi et R. Lafosse : Généralisation de la régression simple pour analyser la dépendance de K ensembles de variables avec un K+1ième. *Revue de statistique appliquée*, XLIX(1):5–30, 2001.

A.E. Hoerl et R.W. Kennard : Ridge regression : biaised estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

P. Horst : Relations among m sets of measures. *Psychometrika*, 26(2):129–149, 1961.

D.W. Hosmer et S. Lemeshow : *Applied logistic regression*. John Wiley and Sons, New York, 1989.

H. Hotelling : Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

A. Höskuldsson : PLS regression methods. *Journal of chemometrics*, 2:211–228, 1988.

A. Höskuldsson et K. Svinning : Modelling of multi-block data. *Journal of chemometrics*, 20:376–385, 2006.

D. Hurnik, I.R. Dohoo, A. Donald et N.P. Robinson : Factor analysis of swine farm management practices on prince edward island. *Preventive Veterinary Medicine*, 20 (1-2):135–146, 1994.

J.K. Johansson : An extension of Wollenberg's redundancy analysis. *Psychometrika*, 46:93–103, 1981.

I.T. Jolliffe : *Principal component analysis*. Springer Verlag, New-York, 1986.

H.F. Kaiser : The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:184–200, 1958.

J.R. Kettenring : Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

G. Kissita : *Les analyses canoniques généralisées avec tableau de référence généralisé : éléments théoriques et appliqués*. Thèse de doctorat, Université Paris Dauphine, 2003.

G. Kissita, P. Cazes, M. Hanafi et R. Lafosse : Deux méthodes d'analyse factorielle du lien entre deux tableaux de variables partitionnés. *Revue de statistique appliquée*, LII(3):73–92, 2004.

M.F. Klein : Etude des facteurs de risques de l'entérocolite épizootique du lapin en engraissement. Thèse vétérinaire, Ecole Nationale Vétérinaire de Nantes, Sept. 2002.

T. Kourti : Multivariate dynamic data modeling for analysis and statistical process control of batch processes, strat-ups and grade transitions. *Journal of Chemometrics*, 16:176–188, 2003.

S.Q. Lafi et J.B. Kaneene : An explanation of the use of principal-components analysis to detect and correct for multicollinearity. *Preventive Veterinary Medicine*, 13:261–275, 1992.

R. Lafosse : Analyse de concordance de deux tableaux : monogamie, simultanéités et découpages. *Revue de statistique appliquée*, XLV(3):45–72, 1997.

R. Lafosse, D. Chessel et M. Hanafi : Analogies de structures des vins de cahors. *In 5ièmes journées agro-industrie et méthodes statistiques*, page 10, INRA Versailles, 1997.

R. Lafosse et M. Hanafi : Concordance d'un tableau avec K tableaux : définition de K+1uples synthétiques. *Revue de statistique appliquée*, XLV(4):111–126, 1997.

L. Lebart, A. Morineau et M. Piron : *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 3ième édition édition, 2000.

F. Lescourret et B. Faye : Stratégie statistique du laboratoire d'écopathologie. *Epidémiologie et santé animale*, 20:103–115, 1991.

A. Lorber, L.E. Wangen et B.R. Kowalski : A theoretical foundation for the PLS algorithm. *Journal of chemometrics*, 1:19–31, 1987.

D.J. Louwerse, A.K. Smilde et H.A.L. Kiers : Cross-validation of multiway component models. *Journal of chemometrics*, 13:491–510, 1999.

F. Madec, N. Bridoux, S. Bounaix et A. Jestin : Measurement of digestive disorders in the piglet at weaning and related risk factors. *Preventive Veterinary Medicine*, 35:53–72, 1998.

F. Madec et C. Fourichon : Les facteurs de risque en épidémiologie animale. *Epidémiologie et santé animale*, 18:31–43, 1990.

F. Madec et J. Josse : Utilisation des méthodes d'analyse des données pour l'étude des maladies d'élevage. Application au porc. *Epidémiologie et santé animale*, 6:35–63, 1984.

F. Madec et J.P. Tillon : Ecopathologie et facteurs de risque en médecine vétérinaire. *Rec. Méd. Vét.*, 164(8-9):607–616, 1988.

T. Manske, J. Hultgren et C. Bergsten : Prevalence and interrelationships of hoof lesions and lameness in swedish dairy cows. *Preventive Veterinary Medicine*, 54 (3):247–263, 2002.

A. Martrenchar, E. Boilletot, D. Huonnic et F. Pol : Risk factors for foot-pad dermatitis in chicken and turkey broilers in France. *Preventive Veterinary Medicine*, 52(3-4):213–226, 2002.

B.D. Marx : Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38(4):374–381, 1996.

W.F. Massy : Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60:234–256, 1965.

R. Meyer : Extension of correspondence analysis for the statistical exploration of multidimensional contingency tables. *In* O.Opitz, éditeur : *Conceptual and numerical analysis of data*, pages 178–186. Berlin, 1989.

E. Morignat, A.G. Biacabe, C. Ducrot, T. Baron et D. Calavas : Typologie des phénotypes biochimiques de l'ESB par analyse discriminante. *Epidémiologie et santé animale*, 49:29–35, 2006.

K.E. Muller : Relationships between redundancy analysis, canonical correlation and multivariate regression. *Psychometrika*, 46(2):139–142, 1981.

J. Obadia : L'analyse en composantes explicatives. *Revue de statistique appliquée*, XXVI(4):5–28, 1978.

D.J. O'Brien, R.H. Poppenga et C.W. Ramm : An exploratory analysis of liver element relationships in a case series of common loons (gavia immer). *Preventive Veterinary Medicine*, 25(1):37–49, 1995.

S.L. Ott, R. Johnson et S. J. Wells : Association between bovine-leukosis virus seroprevalence and herd-level productivity on US dairy farms. *Preventive Veterinary Medicine*, 61(4):249–262, 2003.

J. Pagès et M. Tenenhaus : Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. *Chemometrics and Intelligent Laboratory Systems*, 58(2):261–273, 2001.

R. Palm et A.F. Iemma : Quelques alternatives à la régression classique dans le cas de la colinéarité. *Revue de statistique appliquée*, XLIII(2):5–33, 1995.

J. Pontier et M. Normand : A propos de généralisation de l'analyse canonique. *Revue de statistique appliquée*, XL(1):54–75, 1992.

E.M. Qannari et M. Hanafi : A simple continuum regression approach. *Journal of chemometrics*, 19:387–392, 2005.

S.J. Qin, S. Valle et M.J. Piovoso : On unifying multiblock analysis with application to decentralized process monitoring. *Journal of chemometrics*, 15:715–742, 2001.

C.R. Rao : The use and interpretation of principal component analysis in applied research. *Sankhya, A.*, 26:329–358, 1964.

N. Rose, A. Abherve-Guegen, G. Le Diguerher, E. Eveno, J.P. Jolly, P. Blanchard, A. Oger, A. Jestin et F. Madec : Effet de la génétique piétrain sur l'expression clinique de la maladie de l'amaigrissement du porcelet (MAP). Etude dans quatre élevages naisseurs-engraisseurs. *In Journées de la recherche porcine*, pages 339–344, Paris, 2004.

N. Rose, G. Larour, G. Le Digherher, E. Eveno, J.P. Jolly, P. Blanchard, A. Oger, M. Le Dinma, A. Jestin et F. Madec : Risk factors for porcine post-weaning multisystemic wasting syndrome (PMWS) in 149 french farrow-to-finish herds. *Preventive Veterinary Medicine*, 61:209–225, 2003a.

N. Rose, J.P. Mariani, P. Drouin, J.Y. Toux, V. Rose et P. Colin : A decision-support system for salmonella in broiler-chicken flocks. *Preventive Veterinary Medicine*, 59 (1-2):27–42, 2003b.

R. Rosipal et N. Krämer : Overview and recent advances in partial least squares. *In* C.Saunders & al., éditeur : *Subspace, latent structure and feature selection techniques*, pages 34–51. Springer, 2006.

C. W. Rougoor, W. J. Hanekamp, A. A. Dijkhuizen, M. Nielen et J. B. Wilmink : Relationships between dairy cow mastitis and fertility management and farm performance. *Preventive Veterinary Medicine*, 39(4):247–64, 1999.

R. Sabatier : Analyse factorielle de données structurées et métriques. *Statistique et analyse des données*, 12(3):75–96, 1987.

G. Saporta : *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. Thèse de doctorat, Université Paris VI, 1975.

G. Saporta : *Probabilités, analyse des données et statistique (2nd édition)*. Technip, Paris, 2006.

SAS : *SAS OnLineDoc version 9.1*. SAS Institute Inc., Cary, NC, 2004.

M. Stone : Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36(1):111–147, 1974.

M. Stone et R.J. Brooks : Continuum regression : cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society*, 52 (2):237–269, 1990.

R. Sundberg : Continuum regression and ridge regression. *Journal of the Royal Statistical Society*, 55(3):653–659, 1993.

M. Tenenhaus : *La régression PLS. Théorie et pratique*. Technip, Paris, 1998.

M. Tenenhaus : L'approche PLS. *Revue de statistique appliquée*, 47(2):5–40, 1999.

M. Tenenhaus, J. Pages, L. Ambroisine et C. Guinot : PLS methodology to study relationships between hedonic judgements and product characteristics. *Food quality and preference*, 16:315–325, 2005a.

M. Tenenhaus et V.E. Vinzi : PLS regression, PLS path modelling and generalized procustean analysis : a combined approach for multiblock analysis. *Journal of Chemometrics*, 19:145–153, 2005.

M. Tenenhaus, V.E. Vinzi, Y.M. Chatelin et C. Lauro : PLS path modeling. *Computational statistics and data analysis*, 48:159–205, 2005b.

C.J.F. Ter Braak : Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167–1179, 1986.

M.B. Thoefner, A.K. Ersboll, A.L. Jensen et Hesselholt M. : Factor analysis of the interrelationships between clinical variables in horses with colic. *Preventive Veterinary Medicine*, 48:201–214, 2001.

P.T. Thomsen, S. Ostergaard, H. Houe et J.T. Sorensen : Loser cows in danish dairy herds : Risk factors. *Preventive Veterinary Medicine*, In press, 2007.

J. Tobin : Estimation for relationships with limited dependent variables. *Econometrica*, 26(1):24–36, 1958.

B. Toma, J.J. Benet, B. Dufour, M. Eloit, F. Moutou et M. Sanaa : *Glossaire d'épidémiologie animale*. Editions du Point Vétérinaire, Maisons Alfort, 1991.

B. Toma, B. Dufour, M. Sanaa, J.J. Benet, P. Ellis, F. Moutou et A. Louza : *Epidémiologie appliquée à la lutte collective contre les maladies animales transmissibles majeures*. AEEMA, Maisons Alfort, 1996.

R. Tomassone, E. Lesquoy et C. Millier : *La régression. Nouveaux regards sur une ancienne méthode statistique*. INRA et Masson, Paris, 1983.

L.R. Tucker : An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136, 1958.

J.P. Van de Geer : Linear relations among K sets of variables. *Psychometrika*, 49 (1):79–94, 1984.

A. Van Den Wollenberg : Redundancy analysis : an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219, 1977.

E. Venot : Risk factors for porcine circovirus type 2 (PCV2) positive serology in 159 french farrow-to-finish herds. Rapport de stage, Université de Limburg, 2003.

E. Vigneau, D. Bertrand et E.M. Qannari : Application of latent root regression for calibration in near-infrared spectroscopy. comparison with principal component regression and partial least squares. *Chemometrics and intelligent laboratory systems*, 35:231–238, 1996.

E. Vigneau, E.M. Qannari et D. Bertrand : A new method of regression on latent variables. application to spectral data. *Chemometrics and intelligent laboratory systems*, 63:7–14, 2002.

H.D. Vinod : Canonical ridge and econometrics of joint production. *Journal of econometrics*, 4:147–166, 1976.

M. Vivien : Nouvelles approches en analyse multi-tableaux (rapport de DEA). Rapport technique, Université Montpellier II, 1999.

M. Vivien : *Approches PLS linéaires et non-linéaires pour la modélisation de multi-tableaux : théorie et applications*. Thèse de doctorat, Université de Montpellier 1, 2002.

M. Vivien, T. Verron et R. Sabatier : Comparing and predicting sensory profiles by NIRS : Use of the GOMCIA and GOMCIA-PLS multi-block methods. *Journal of chemometrics*, 19:162–170, 2005.

B.A. Wagner, M.D. Salman, D.A. Dargatz, P.S. Morley, T.E. Wittum et T.J. Keefe : Factor analysis of minimum-inhibitory concentrations for Escherichia coli isolated from feedlot cattle to model relationships among antimicrobial-resistance outcomes. *Preventive Veterinary Medicine*, 57(3):127–139, 2003.

L.E. Wangen et B.R. Kowalski : A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of chemometrics*, 3:3–20, 1988.

T. Webster, R.F. Gunst et R.L. Mason : Latent root regression analysis. *Technometrics*, 16(4):513–522, 1974.

L.A. Weissfeld et S.M. Sereika : A multicollinearity diagnostic for generalized linear models. *Commun. Statist. - Theory Meth.*, 20(4):1183–1198, 1991.

J.A. Westerhuis et P.M.J. Coenegracht : Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of chemometrics*, 11(5):379–392, 1997.

J.A. Westerhuis, T. Kourti et J.F. MacGregor : Analysis of multiblock and hierarchical PCA and PLS model. *Journal of chemometrics*, 12:301–321, 1998.

J.A. Westerhuis et A.K. Smilde : Deflation in multiblock PLS (short communication). *Journal of chemometrics*, 15:485–493, 2001.

B.M. Wise et N.L. Ricker : Identification of finite impulse response models with continuum regression. *Journal of Chemometrics*, 7(1):1–14, 1993.

H. Wold : Estimation of principal components and related models by iterative least squares. *In* Krishnaiah, éditeur : *Multivariate analysis*, pages 391–420. Academic press, New York, 1966.

H. Wold : Soft modelling : the basic design and some extensions. *In* K.G Jöreskog et H. Wold, éditeurs : *System under indirect observation. Part* 2, pages 1–54. North-Holland, Amsterdam, 1982.

H. Wold, P. Geladi, K. Esbensen et J. Ohman : Multi-way principal components and PLS analysis. *Journal of chemometrics*, 1:41–56, 1987.

S. Wold : Three PLS algorithms according to SW. *In* S. Wold, éditeur : *Symposium MULDAST (multivariate analysis in science and technology)*, pages 26–30, Umea University, Sweden, 1984.

S. Wold, N. Kettaneh et K. Tjessem : Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics*, 10:463–482, 1996.

S. Wold, H. Martens et H. Wold : The multivariate calibration problem in chemistry solved by the PLS method. *In* Ruhe A. B. et Kastrom, éditeurs : *Proceedings of the Conference on Matrix Pencils*, pages 286–293. Springer Verlag, Heidelberg, 1983.

P.S.A. Woods, H.J. Wynne, H.W. Ploeger et D.K. Leonard : Path analysis of subsistence farmers'use of veterinary services in Zimbabwe. *Preventive Veterinary Medicine*, 61:339–358, 2003.

S. Wright : Correlation and causation. *Jour. Agric. Res.*, 20:557–585, 1921.

Index

Résumé

Ce travail de recherche s'inscrit dans le cadre des méthodes factorielles qui permettent de décrire et prédire des données structurées en plusieurs tableaux. Les objectifs et la nature des données d'épidémiologie analytique dans le domaine vétérinaire ont amené à centrer le travail sur les méthodes de régression multibloc, qui orientent la description de plusieurs tableaux de variables vers l'explication d'un autre tableau. Un des principaux objectifs est de contribuer à la réflexion sur la sensibilité de ces méthodes à la multicolinéarité. Des méthodes statistiques existantes sont présentées et reliées dans un cadre unifié, relevant soit de critères à maximiser comparables, soit d'un continuum général les reliant. De nouvelles méthodes peu vulnérables à l'égard de la multicolinéarité, et s'appliquant au cas de données structurées en deux puis en $(K+1)$ tableaux, sont proposées. L'intérêt de ces méthodes, ainsi que des continuums qui leur sont associés, est illustré sur la base d'études de cas réels en épidémiologie. Ce travail de recherche a permis d'appliquer les méthodes multiblocs au domaine de l'épidémiologie animale, dans lequel elles n'avaient pas encore été utilisées.

Mots-Clés : *Analyse factorielle, régression multibloc, multicolinéarité, analyse en composantes principales sur variables instrumentales, régression Partial Least Square, Latent Root Regression, approche continuum.*

Abstract

This work deals with factor analysis method to investigate relationships among several data sets in veterinary epidemiology. Considering the aims and the nature of data collected within this framework, the research work is focused on the multiblock setting for the purpose of exploring and modelling one data set from several other data sets. The key feature is to handle the sensitivity of these methods to multicolinearity problem among the explanatory variables. Existing methods are described and unified by considering the criteria to be optimised or by setting up a continuum approach. New formulations of alternative methods which are robust to the multicolinearity problem are discussed. These methods can be applied to data organised in two or several data sets. The interest of the general strategy of analysis, related to the methods and continuums, is illustrated on the basis of real case studies pertaining to veterinary epidemiology.

Keywords : *Factor analysis, multiblock regression, multicolinearity, principal component analysis on instrumental variables, Partial Least Square regression, Latent Root Regression, continuum approach.*